Antigen presentation and the boundary between self and non-self: applications in computational immunology

Thesis submitted for the degree of Doctor of Philosophy

Antigen presentation and the boundary between self and non-self: applications in computational immunology

by

Nazmiye Uzun

Birkbeck, University of London

13th May, 2022

I, Nazmiye Uzun, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

This thesis focuses on the applications of computational prediction methods in the areas of antigen presentation and recognition close to the boundary between self and non-self. Publicly available MHC binding prediction tools are combined with data about the frequencies of different HLA alleles within global and regional populations. Additionally, a new *in silico* method called proteome scanning is presented that assesses whether, in the light of central tolerance mechanisms that remove self-reactive T cells, an individual is likely to have T cells capable of binding to a given peptide-MHC surface.

These methods are applied in three main application areas. Firstly, predictions are made concerning individuals with missense mutation haemophilia A and their risk of developing inhibitors against the replacement Factor VIII used to treat the condition. Inhibitor formation is known to be a $CD4^+$ T cell-dependent process; the analysis presented here demonstrates that understanding the risk of inhibitor formation generally requires knowledge of an individual's HLA types.

Secondly, predictions are made concerning the risk of transplant rejection and suggest that the proteome scanning approach can be used to predict whether a given HLA mismatch between donor and recipient is likely to increase rejection risk.

Thirdly, predictions are made concerning candidate peptide biomarkers (derived from known tumour antigens) for hepatocellular carcinoma, with the selection of peptide pools optimised in terms of MHC binding affinity and global population coverage. These are currently undergoing laboratory evaluation at the Institute of Hepatology, with promising early results.

Taken together, these applications illustrate the breadth of potential contributions to clinical practice afforded by the computational prediction of antigen presentation and recognition close to the self/non-self boundary, including the novel proteome scanning methodology.

Acknowledgements

I would like to express my gratitude to my supervisor, mentor and teacher Prof. Adrian Shepherd for his support, invaluable advice and assistance throughout my masters and doctoral studies. I cannot thank him enough for being the guiding light for the path I have chosen in my academic research; I am deeply grateful for his time and patience, and the immense knowledge that he has shared with me.

I would like to extend my gratitude to my second supervisor, the director of the Institute of Hepatology, Dr. Shilpa Chokshi. I really appreciate her making it possible to be part of a highly promising and pivotal project, that will perhaps soon enable us to have an impact on the future of cancer patients.

I would also like to thank Dr. Irilenia Nobeli for her support all along during my journey at Birkbeck. I am most grateful for her insightful comments and suggestions, constructive feedback and her warm encouraging smile.

I owe special thanks to Prof. David Moss, the enlightening voice that guided our minds when we needed it the most, providing his vast experience and extensive knowledge across diverse subjects.

Many thanks to Bora, my other half, for his never-ending support and understanding at every step. Last, but not the least, thanks to Sumer, my little daughter, the inspiration that is constantly leading me to strive to contribute for a better future for her and for the whole world.

Table of Contents

List of Figures

List of Tables

1 Introduction

The focus of this thesis is on diverse applications of computationally-predicted antigen presentation and recognition within human hosts. This introductory chapter covers the background knowledge and concepts relevant to all of these applications: the role and mechanisms of antigen presentation (section 1.1) and recognition (section 1.2) within the T cell branch of the adaptive immune system. The following chapter covers key computational resources and prediction methods within this area (chapter 2). Background information relevant to specific applications is deferred until subsequent chapters: alloimmune responses to replacement protein therapeutics (chapter 3); alloimmune responses in the context of organ transplantation (chapter 4); and the identification of broadly-applicable cancer biomarkers (chapter 5).

1.1 Antigen presentation

Major histocompatibility complex (MHC) molecules are transmembrane proteins that present peptides to T cells through the formation of peptide-MHC complexes. This process, which involves both self- and non-self-peptides, is known as *antigen presentation* and takes place on the surface of host cells. A presented peptide that is recognised by a T cell receptor (TCR) is called a *T cell epitope*.

There are two contrasting antigen presentation pathways involving different types of MHC molecules: MHC class I and MHC class II molecules. MHC class I molecules are expressed by all nucleated cells and present peptide fragments of endogenously synthesised proteins to $CD8^+$ (cytotoxic) T cells. $CD8^+$ T cells represent a major subset of T cells capable of killing host cells that are infected with intracellular pathogens, notably viruses together with certain species of bacteria (e.g., *Listeria monocytogenes* and *Chlamydia trachomatis*) and protozoans (including members of the malaria-causing genus *Plasmodium*). (The nomenclature $CD8^+$ refers to the presence of the CD8 molecule on the surface of these cells and is part of a widely-used protocol for differentiating between different subsets of immune cells [Engel et al., 2015].)

MHC class II molecules, on the other hand, are predominantly expressed by professional antigen presenting cells (APCs), which internalise antigens of exogenous origin. APCs — notably dendritic cells, macrophages and mature B cells — present peptide fragments to

$CD4^+$ (helper) T cells via MHC class II molecules. $CD4^+$ T cells, a second major subset of T cells, are associated with various roles — they have been described as the "orchestrators, regulators and direct effectors of antiviral immunity" (Swain et al., 2012) — that includes helping B cells to generate a mature antibody response.

In this section, the focus will be on practical issues of relevance to the research undertaken for this thesis, while also engaging with underlying biological mechanisms: What lengths of peptides are presented and what sequence preferences do they have? What is the conformation of bound peptides? And what determines the relative frequencies of different presented peptides (known collectively as the immunopeptidome [Vizcaíno et al., 2020]) and how long do they remain bound to MHC?

1.1.1 MHC class I presentation pathway

Unwanted or damaged proteins are broken down in the cytoplasm or nucleus of cells by proteasomes. Eukaryote proteasomes are large protein complexes that play a vital role in several fundamental processes (for a useful overview, see [Marques et al., 2009]), but here the focus is the contribution to antigen presentation. Under immune stress and interferon-γ (IFN-γ) stimulation, synthesis of "standard" proteasome complexes (known as constitutive proteasomes) switches to immunoproteasome synthesis, involving the replacement of three catalytically active subunits (Basler et al., 2013). The catalytic activity of both constitutive proteasomes and immunoproteasomes generates short peptides in the range 3 to 22 residues (Kisselev et al., 1999), whereas most peptides bound to MHC class I (MHC-I) molecules are of length 8 to 10 residues, constrained by the length of the MHC-I groove, which is blocked at both ends (**Figure 1.1**). The pool of MHC-I-compatible peptides is increased by post-proteasomal N-terminal trimming (Rock et al., 2004). Peptides of length 8 to 16 residues are preferentially transported to the endoplasmic reticulum (ER) by the transporter associated antigen processing (TAP) protein complex, where suitable peptides are loaded onto MHC-I molecules by the peptide-loading complex (Cresswell et al., 1999), after which the peptide-MHC complex is transported to the cell surface by the secretory pathway.

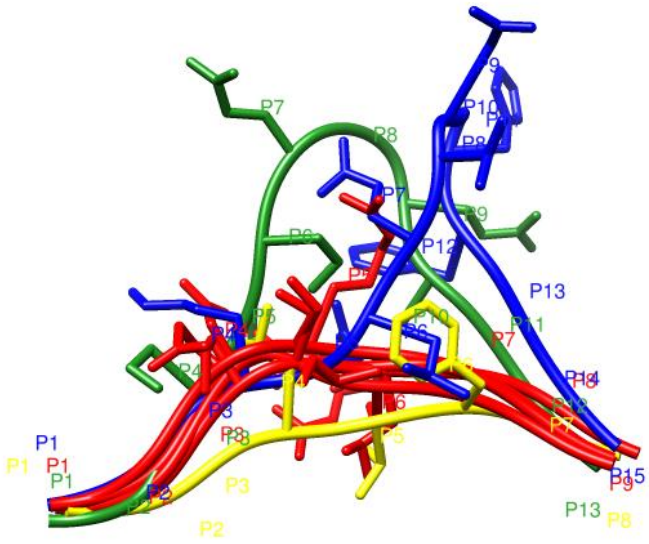


Figure 1.1 Comparison of variable length peptides accommodated in the groove of MHC-I molecules. The peptide backbone structures are demonstrated in tube representation and side chains are in stick form. The peptides were removed from their bound MHC molecules, superimposed to align, and arrange so that their TCR facing residues point upwards. The 8-mer peptide, isolated from the structure 5HGB, is coloured in yellow; the 9-mers, isolated from 3RL1, 5SWQ, 7JYV and 7KGQ, are coloured in red. The 13mer peptide from 2AK4 and 15mer from 4U6Y, which are atypically long for an MHC-I groove closed on both termini, are coloured in green and blue, respectively. Figure inspired by Rudolph et al (2006).

Several steps of the pathway contribute to the shape of the MHC-I immunopeptidome — the pool of peptide that gets presented to $CD8^+$ T cells by MHC-I molecules. Certain peptides are preferentially cleaved by the proteasome, with each catalytic subunit having its own, distinct substrate specificity. These distinct specificities are reflected in differences between the immunopeptidome associated with constitutive proteasomes and that associated with immunoproteasomes, although a recent in-depth study (using a mass spectrometry-based strategy described as both "global" and "unbiased") contradicts some earlier findings (e.g. those of [Toes et al., 2001]) and concluded that, although the frequency of certain MHC-I epitopes is different, the immunoproteasome pool "does not appear to be preferentially suited for antigen presentation" (Winter et al., 2017). ERAP1, the key aminopeptidases associated with N-terminal trimming, has been observed to have strong amino-acid preferences at certain positions, including hydrophobic amino acids at the peptide's C-terminus (Evnouchidou et al., 2008). Prior to the transportation of the peptide-MHC complex to the cell surface, the chaperone tapasin, a component of the peptide-loading complex, promotes the selection of peptides that bind to MHC-I molecules with slow off rates.

But arguably the most selective step is the binding of the peptide to the MHC-I molecule, discussed below in section 1.1.3.

Although the preceding account summarises the orthodox view of MHC class I antigen presentation, there are two important areas of controversy. Firstly, it is known that proteases other than the proteasome are capable of generating MHC-I peptides, and some have gone so far as to argue that the importance of the proteasome may have been exaggerated (Milner et al., 2013). Secondly, it is known that some MHC-I peptides are created by splicing as well as cleavage, although the proportion of peptides within the MHC-I immunopeptidome is hotly disputed, with some arguing that it is close to zero and others that it is potentially as high as 45% (Purcell, 2021). Whereas the former point (i.e., the mode of cleavage) is mechanistically important, the latter point (i.e., splicing frequency) is of profound conceptual importance, as it potentially transforms our understanding of the boundary between self and non-self. However, although this dispute has been around since at least 2016 (Liepe et al., 2016; Mylonas et al., 2018), it is currently far from settled (Purcell, 2021).

1.1.2 MHC class II presentation pathway

The MHC class II (MHC-II) pathway is mainly associated with professional APCs, although MHC-II expression by various other cell types can be induced (for example by IFN-γ) (Neefjes et al., 2011). MHC-II molecules are assembled in the ER and a special peptide called CLIP (Class II-associated invariant chain peptide) binds to its groove, but subsequently displaced by an antigenic peptide with a higher binding affinity. Such antigenic peptides derive from internalised proteins that are broken down in endosomal or lysosomal compartments by proteases, most notably by members of the cathepsin group. Different cathepsins are expressed at different levels in different professional APCs (Hsing & Rudensky, 2005).

A useful summary of the rather complicated mechanisms underpinning MHC-II formation, loading and transport to the cell surface can be found in (Neefjes et al., 2011). The main topic of interest here concerns the characteristics of the peptides available for MHC-II binding and presentation. Whereas the groove of MHC-I molecules is blocked at both ends, that of MHC-II molecules is open-ended and can accommodate much longer peptides. Lengths ranging between 10 and 34 residues have been observed (Chicz

et al., 1993), with the same 9-mer core sub-sequence often recurring in multiple sequences of different lengths, known as a nested set (Lippolis et al., 2002). Peptide length has a significant impact on binding affinity, with lengths of around 18-20 residues considered optimal (O'Brien et al., 2008), and the presence or absence of specific flanking residues may impact $CD4^+$ T cell function (Holland et al., 2013). Although the cleavage preferences of certain cathepsins have been investigated experimentally (O'Donoghue et al., 2012), our understanding of the factors that shape the pool of peptides available for binding to MHC-II molecules is far from complete.

It is also worth noting that professional APCs, most notably dendritic cells, are capable of presenting extracellular antigens via MHC-I molecules, a process known as cross-presentation. Two main pathways have been observed: the vacuolar pathway that involves antigen degradation by cathepsins and the endosome-to-cytosol pathway that involves proteasomal cleavage (Embgenbroich & Burgdorf, 2018).

1.1.3 MHC structure and binding

MHC molecules play a critical role in both antigen presentation pathways and deserve detailed consideration in their own right. MHC molecules of both class I and class II share a similar structural form. They are both heterodimers with transmembrane helices (one in MHC-I, two in MHC-II) that secure the molecule on the cell surface. Both have a pair of α-helices forming the sides of the peptide-binding groove and a seven-stranded β-sheet forming its floor, and the floor has binding pockets that accommodate the side-chains of amino acids from the peptide (Rudolph et al., 2006). **Figure 1.2** shows structural units that form MHC-I and MHC-II molecules with their bound peptides and interacting TCRs.

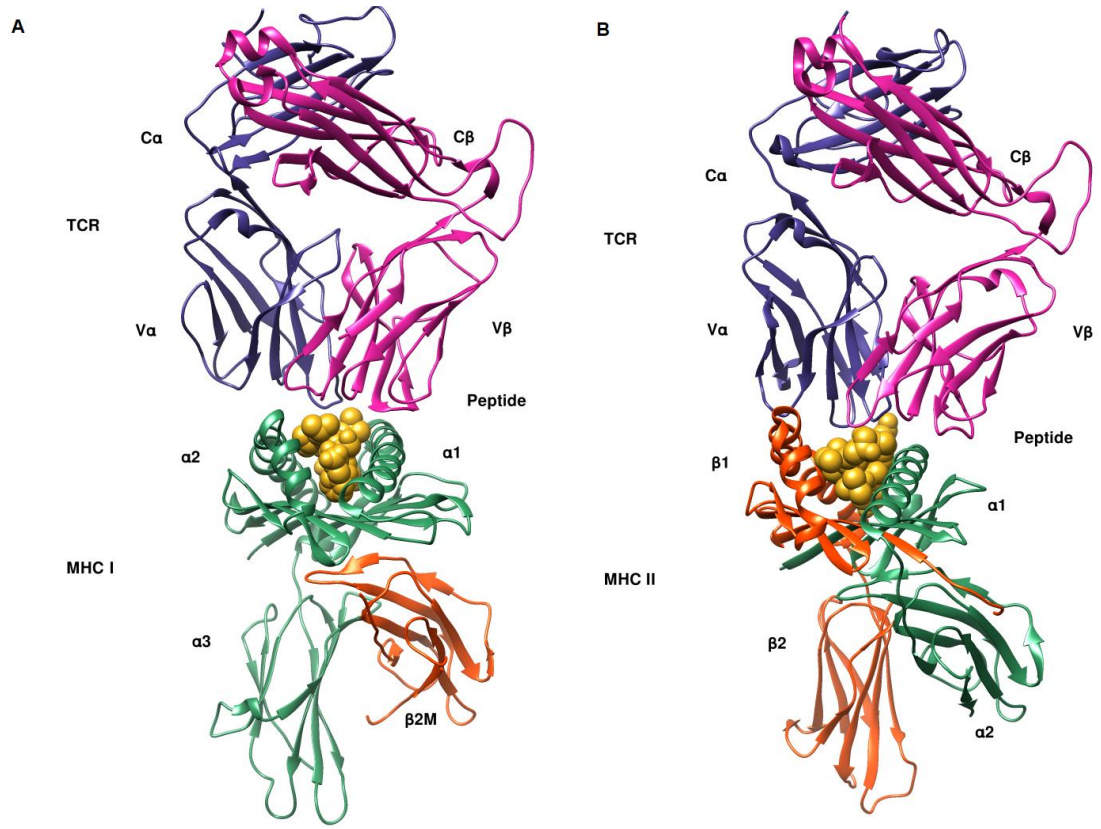


Figure 1.2 Structures of an MHC I-peptide-CD8$^+$ T cell complex (PDB accession code 6MTM) and an MHC II-peptide-CD4$^+$ T cell complex (PDB accession code 6R0E) in image A and B respectively. The TCR constant (C) and variable (V) regions are displayed with α and β chain annotation. The peptide is displayed in yellow. MHC-I α helices, α1 and α2, that form the binding groove, and the α3 domain are in green. The β2M domain is coloured in orange. MHC-II α domains are in green and the β domains are coloured in orange.

At the same time, there are several differences between the MHC-I and MHC-II molecules. In the MHC-I heterodimer, the groove is formed by two copies of the α chain (domains α1 and α2), and domain α3 is involved in CD8 coreceptor recognition, while the β chain is small and invariant. In the MHC-II heterodimer, the β chain is only moderately shorter than the α chain, both are polymorphic, both contribute to the formation of the MHC-II groove (domains α1 and β1), and both contribute to CD4 co-receptor recognition (domains α2 and β2). Most importantly from an antigen presentation perspective, the MHC-I and MHC-II grooves are different. The MHC-I groove is blocked at both ends, and typically has binding pockets at positions P1 and P9, whereas the MHC-II groove is open at both ends and typically has pockets at positions P1, P4, P6 and P9 (Rudolph et al., 2006).

A key event in both presentation pathways is the binding of peptides within the MHC groove.

The contrast between blocked and open grooves is the major factor determining the length differences between class I and class II peptides, as discussed in section 1.1.1. However, it is worth noting that, although the close groove of MHC-I commonly restricts binding to peptides of length 8 to 10, it has been recognised for some time that longer peptides — up to at least 14 residues — can bind to MHC-I molecules with high affinity (see **Figure 1.1**) and "elicit dominant cytotoxic T lymphocyte responses" (Burrows et al., 2006). Owing to their "super-bulging" conformation, these longer peptides can adopt unusual conformations and form diverse interactions with the TCR. Indeed, a recent analysis of 29 structural complexes involving MHC-I molecules and long antigenic peptides found that the bulged peptides were capable of forming rigid secondary structures such as β-hairpins, β-turns or α-helices, and these facilitated the formation of more diverse TCR docking modes, some of which involve unusually large numbers of contacts with the peptide and correspondingly few contacts with the MHC molecule (Josephs et al., 2017). (Canonical TCR-peptide-MHC binding is discussed in section 1.2.)

Not all pairings of MHC molecule and peptide have equal potential to be immunogenic (i.e., induce a T cell response). The most widely used correlate of immunogenicity is binding affinity. Competition binding assays indicate that around 80% of class I epitopes have an of $IC_{50} \leq 500$ nmol/L (Sette et al., 1994; Paul et al., 2013), whereas a threshold of $IC_{50} \leq 1{,}000$ nmol/L is commonly adopted for class II epitopes (Southwood et al., 1998; Paul et al., 2015; Paul et al., 2020). In both cases, the distribution of binding affinities varies between alleles (the allelic variability of the genes that encode MHC molecules is discussed in section 1.1.4) (Paul et al., 2013; Paul et al., 2020), which has potential implications for the choice of thresholds for deciding whether a given MHC-binding peptide is a potential epitope.

Although binding affinity is widely used, it has long been argued that immunogenicity is better correlated with the stability of the peptide-MHC (pMHC) complex than with its affinity — originally in the context of MHC-I (van der Burg et al., 1996) and more recently MHC-II (Lazarski et al., 2005). The rationale is straightforward enough: the "sustained presentation" of a peptide increases the chances that a TCR will encounter it, and this implies "some degree" of peptide-MHC stability (Harndahl et al., 2012). However,

although assays for measuring the rate of peptide dissociation have been developed for MHC-I (Harndahl et al., 2011) and MHC-II (Chaves & Sant, 2007), they do not appear to be widely used.

1.1.4 The genetics of MHC molecules

An important role of MHC molecules is to ensure that a variety of peptides are presented at both the individual and population levels so that it is much harder for pathogens to successfully evade presentation by mutating. In this context, the genes that encode MHC molecules exhibit two key properties: they are polygenic (i.e., each individual has multiple class I and class II genes and therefore sets of different MHC molecules within each class); and they are highly polymorphic (key MHC genes are highly variable at the population level) (Janeway et al., 2001).

Human MHC molecules are encoded by the Human Leukocyte Antigen (HLA) region of the human genome located on chromosome 6 at position 6p21.3 and consists of around 4 megabases. The HLA region contains genes associated with a variety of functions, not all of them immune-related (for a thorough survey of the region, see [Shiina et al., 2009]), but the focus here is exclusively on MHC. Humans have multiple class I and class II MHC molecules. These can be broadly divided into "classical" and "non-classical" MHC molecules. The former are highly polymorphic, have been extensively studied, and represent a major aspect of the research presented in this thesis; the latter are monomorphic (or nearly monomorphic), are often associated with the surveillance of entities other than standard peptides (e.g., lipids and post-translationally modified peptides), and have arguably been neglected (D'Souza et al., 2019). The focus of this research has been on a subset of classical MHC molecules.

There are 3 classical HLA loci that encode for MHC class I molecules — HLA-A, -B and -C — and 3 classical loci that encode for MHC class II molecules — HLA-DP, -DQ and -DR. In class II molecules, both α and β chains contribute to peptide binding and are polymorphic in HLA-DQ and -DR (though not -DP), hence the letters A and B are commonly appended (e.g., HLA-DQA and HLA-DQB). HLA-DR β chains are encoded by 4 loci: HLA-DRB1, -DRB3, -DRB4 and -DRB5. HLA-DRB1 is expressed in all haplotypes, with one of DRBs 3, 4, 5 present on each chromosome (hence any individual will have at most 3 of the 4 HLA-DR molecules). Linkage disequilibrium between certain

HLA loci is well established, although a comprehensive picture has yet to emerge. One example is that between theDRB1 genes and the DRB3/4/5 genes (Dorak et al., 2002).

MHC encoding genes are polygenic and among the most polymorphic in human genome. As of 24th April 2022, the IPD-IMGT/HLA database reports that there are more than 24,000 HLA class I alleles and 9,000 HLA class II alleles (Robinson et al., 2020). MHC molecules are implicated to some extent in human mate selection via body odour (Dandine-Roulland et al., 2019); mating between individuals with different MHC molecules increases the likelihood that the offspring will be heterozygous with respect to their MHC molecules. Given that the HLA allelic variations are disproportionately associated with the encoding of residues that contribute to the peptide binding groove, MHC heterozygosity may increase an individual's chances of combating a wider range of pathogens.

MHC polymorphism may also be beneficial at the population level, as it increases the chances that some individuals will have the "ideal" HLA haplotypes for supporting an effective and strong immune response against a given infective agent. In accordance with this view, certain HLA alleles and haplotypes have been associated with susceptibility to, or protection against, various diseases. Hodgkin's lymphoma was the first disease found to be associated with HLA-B (Amiel, 1967). Since then, HLA has been associated with many autoimmune and infectious diseases. A meta-analysis study including 42 genome-wide linkage studies for 11 autoimmune diseases determined that the most significant link obtained was with the HLA region (Forabosco et al., 2009).

1.2 Antigen recognition

Section 1.1 explained how a complex is formed between an MHC molecule and an antigenic peptide (commonly derived from a pathogen or a self-protein). The surface formed by this complex is the target for T cell receptors (TCRs) on the surface of T cells, which play a critical role in the adaptive immune response, including killing infected host cells ($CD8^+$ T cells) and regulating the B cell (antibody) response ($CD4^+$ T cells). Whereas antigen presentation makes no distinction between peptides of self and non-self origin, it is crucial that such a distinction is made if the adaptive immune system is to mount a vigorous response against foreign antigens but not an autoimmune response that targets self-proteins. The role of "policing" the boundary between self and non-self rests

primarily with T cells, and depends on complex T cell development and selection mechanisms that originate in the thymus.

Each TCR consists of two amino-acid chains. Most T cells are known as αβ T cells with TCRs consisting of a single α and a single β chain. A subset of T cells is γδ T cells (with γ and δ chains) that (though poorly understood) appear to be associated with lipid antigens and to lie outside the "classical" antigen presentation pathways of αβ T cells (Adams et al., 2015). Further consideration of γδ T cells lies outside the scope of this thesis.

The remainder of this section will focus on practical issues of relevance to the research presented in the thesis, while also engaging with some of the broader characteristics of the T cell response, and briefly with some of the underlying mechanisms: How are responses to self-peptides (i.e., autoimmune responses) generally prevented? How diverse are T cells within a human repertoire? How do TCRs bind to peptide-MHC (pMHC) complexes? To what extent does the response of an individual to given challenge focus on one or more antigenic peptides rather than more broadly across a larger number of epitopes? And to what extent are responses to the same challenge shared between individuals?

1.2.1 Thymic antigen presentation and the self-/non-self boundary

New T cells are produced by haematopoiesis within the bone marrow, but subsequently migrate to the thymus. Immature T cells in the thymus are known as thymocytes, and each has a single type of TCR, but within the population of thymocytes, TCR sequences are highly diverse (as discussed in section 1.2.2). However, this initial diversity is reduced by selection processes that depend on their exposure to complexes between a self-peptide and an MHC molecule (self-pMHC). The aim here is two-fold: to favour, through *positive selection*, TCRs that may potentially respond to some unknown foreign antigen by weeding out TCRs that are non-responsive to self-pMHC; and to remove, through *negative selection* (also known as *central tolerance*), TCRs that bind strongly to self-pMHC and hence T cells that may potentially lead to autoimmune diseases (Janeway et al., 2001).

Various mechanisms — including so-called "promiscuous gene expression" by medullary thymic epithelial cells (mTECs), and the presentation of blood-born and tissue-specific

antigens by dendritic cells subsets (Hasegawa & Matsumoto, 2018) — combine, in the words of Derbinski and Kyewski, to ensure "a maximal representation of the 'immunological self'" (Derbinski & Kyewski, 2010). Nevertheless, "holes" in central tolerance may occur, for example when certain antigens are expressed at very low levels by mTECs (Klein et al., 2014), and have been linked to autoimmune disease in specific cases (see, for example, Lv, et al., 2011). Self-reactive T cells that escape central tolerance may be picked up by additional mechanisms associated with peripheral tolerance. (For a recent review of distinct "tolerance checkpoints" that contribute to peripheral tolerance, see [EITanbouly & Noelle, 2021].)

One important mechanism contributing to peripheral tolerance involves the suppression of the immune response by regulatory T cells (Tregs). Whereas T cells that bind strongly to self-pMHC are destroyed and those that bind weakly to self-pMHC become (after exposure to their cognate antigens) effector or memory T cells, T cells with an intermediate level of binding become Tregs (Li & Rudensky, 2016). The majority of Tregs are associated with the MHC class II antigen presentation pathway. Given the focus on epitope recognition in the research undertaken for this thesis, one intriguing aspect of Treg biology is the possibility that they recognise epitopes – termed Tregitopes – that have distinctive sequence properties. This perspective has been pioneered by de Groot and colleagues for many years (de Groot et al., 2008), but the prevalence of Tregitopes in proteins other than immunoglobulins is unclear (Cousens et al., 2013), and there are no public methods for Tregitope detection.

Distinguishing non-self from self is a particular challenge for the immune system in cases where non-self is very similar to self, i.e., the number of protein residues that are different in non-self compared to self is small (potential only a single residue). On the one hand, the non-self-specific residue(s) may go undetected, either because they are never presented at TCR-facing positions by the individual's MHC molecules, or because TCRs capable of binding to the corresponding peptide-MHC complex do not exist in sufficient numbers. On the other hand, the proliferation of TCRs that bind to a non-self-epitope that is very similar to a self-epitope may increase the risk of a cross-reactive autoimmune response to that self-epitope and hence the failure of central tolerance. A prerequisite to gaining insights into these effects, it is important to understand the number and position of non-self-specific residue(s), which implies a comparison of non-self and self, where self corresponds to the entire host proteome. In this context, the notable contribution of this

research is that a methodology (described in detail in section 3.2) has been developed that takes into account all the proteins within the host proteome, the predicted MHC binding characteristics of self and non-self peptides, and the likely orientation of residue side-chains with respect to TCR binding in order to assess whether a given non-self epitope is likely to be distinguishable from self.

For the practical purposes of the research presented here, the working assumption is that tolerance to self-peptides is effective — which, in all but a small fraction of cases, it is — and hence, from the perspective of T cell recognition, foreign peptides that are indistinguishable from self-peptides will not induce a T cell response.

1.2.2 T cell receptor diversity

The TCRs found within the T cell repertoire of a single individual are highly diverse. The key underlying process that generates TCR diversity is *V(D)J recombination*. The TCR α chain is formed by the combination of germline gene segments located on chromosome 14q11: a constant (C) segment, one of 44 variable (V) segments, and one of 61 junction (J) segments. The TCR β chain is formed by the combination of germline gene segments located on chromosome 7q34: one of two C segments, one of 64 V segments, one of two diversity (D) segments, and one of 14 J segments. Note, however, that the precise number of genes may vary between individuals owing to duplication and deletion events, although the prevalence of such events is poorly understood in the TCR-encoding loci (Collins et al., 2020).

The diversity attributable to the combination of gene segments, known as *combinatorial diversity*, is further enhanced by the additional and removal of nucleotides at the junction between segments, a process known as *junctional diversity*. The consequence is that the greatest diversity is at the junction of the V, D and J segments of the β chain, which corresponds to the CDR3 loop (the third Complementarity Determining Region) that has a special role in antigen binding (see section 1.2.3). An additional contribution to TCR diversity comes from the pairing of α and β chains.

A recent review paper conveniently summarises some key finding regarding the diversity of αβ TCRs within the T cell repertoire of a single individual (Davis & Boyd, 2019). There are, in effect, two ways of defining an upper bound on TCR diversity: a theoretical

limit derived by pairing all possible α chain sequences with all possible β chain sequences equates to around 10^{15} TCRs, whereas the total number of T cells within a single individual is around 2×10^{11}. But the actual number of unique TCRs is likely much smaller. For example, a 2009 next generation sequencing study based on peripheral blood from two adult donors estimated the number of unique TCR β chains within a single individual to be approximately 3×10^{6} (Robins et al., 2009).

It is worth noting that a 2016 deep sequencing study (not cited by [Davis & Boyd, 2019]) based on blood samples from four infant donors estimated that the thymus contains 40 to 70×10^{6} unique TCR β chains and 60 to 100×10^{6} unique TCR α chains (Vanhanen et al., 2016). Given that these estimates are an order of magnitude higher than those from peripheral blood samples, it appears likely that this study has captured at least part of the TCR diversity exhibited by thymocytes prior to negative selection. However, the reason for the higher reported diversity in the α chain compared to the β chain is unclear and, given the known challenges of estimating TCR diversity from relatively small samples (Laydon et al., 2015), it is worth treating all such estimates with a degree of caution.

In addition to the diversity within individuals, there is diversity at the population level that builds on polymorphisms within the loci that encode TCRs. A recent inferential analysis of TCR repertoire data suggests there are many as yet undocumented germline gene polymorphisms, notably within the TCR V β gene, some of which are "strongly associated with dramatic changes in the expressed repertoire" (Omer et al., 2022). Research into this interesting topic is in its infancy, and the potential implications for the research presented here are currently hard to judge.

1.2.3 The binding of TCRs to peptide-MHC complexes

Each of the two TCR chains have three loops, known as *complementarity determining regions* (CDRs), within their respective V regions that contribute to the binding of a TCR to a given pMHC complex. The position and orientation of binding is notably constrained. Given a vector pointing along the long axis of the MHC groove and a second vector separating the CDRs of the TCR α from those of the TCR β chains, the "crossing angle" between these two vectors, and hence the docking orientation of TCR to MHC molecule, is a diagonal (a precise algorithm for calculating this angle is given in [Rudolph et al., 2006]). The canonical docking orientation ensures that the main contacts between

TCR and antigenic peptide involve the CDR3 loops, with that of the TCR α chain positioned over the N-terminal region of the peptide and that of the TCR β chain positioned over the peptide's C-terminal region. The other CDRs are mainly involved in forming contacts with the MHC molecule (Rossjohn et al., 2015).

As noted in a recent review, crossing angles are commonly in the range 22 to 69 degrees (Barbosa et al., 2021). However, it is worth noting that some dramatically different binding orientations have been observed, such as TCRs that bind with "reversed polarity" (i.e., where the crossing angle is approximately 180 degrees away from the canonical angle) (Beringer et al., 2015). The prevalence of non-canonical binding modes within a "typical" developing T cell response is unclear, as is their relevance in the context of combatting or susceptibility to disease.

Although the number of unique TCRs in a T cell repertoire is very large (section 1.2.2), this number is much smaller than the potential number of potential foreign peptides, estimated to be $>10^{15}$ (assuming peptides of length 8 to 14 residues in length have a 1% to 3% MHC binding rate) (Sewell, 2012). Consequently, it has been argued that, in order to provide sufficient coverage of foreign peptides, it is crucial that TCRs are cross-reactive (or promiscuous), i.e., bind to multiple peptide-MHC surfaces (Sewell, 2012). In certain cases, the degree of cross-reactivity is extreme — witness the paper *A Single Autoimmune T Cell Receptor Recognizes More Than a Million Different Peptides*, in which a large sample of peptides were experimentally verified to bind to the chosen $CD8^+$ TCR with sufficient strength for the association to be considered functionally relevant (Wooldridge et al., 2012). This included a peptide that differed from the original, "reference" peptide (derived from preproinsulin) at 7 out of 10 positions.

Several complementary mechanisms underpin the cross-reactivity of TCRs (Barbosa et al., 2021). Potential differences in the crossing angle between TCR and peptide-MHC complex have already been mentioned, but this may be combined with difference in the vertical angle (tilt) between TCR and MHC (Rudolph et al., 2006). Conformational flexibility is another key factor — in particular, that of the β chain CDR3, in which large shifts in the bound versus unbound conformation have been observed, with at least one known example in excess of 11 Å (Petrova et al., 2012). Certain TCR binding modes are thought to have a particular association with autoimmunity, such as: binding that involves a large shift of the TCR towards the N-terminus of the peptide such that interactions with

the central region of the peptide are minimal or absent; and binding with large tilt that reduces the number CDRs in contact with the peptide (Yin et al., 2012).

Although the mechanisms associated with cross-reactivity are interesting (and discussed more fully in [Barbosa et al., 2021]), their relative prevalence within a typical T cell repertoire is poorly understood.

1.2.4 Systemic properties of T cell responses: immunodominance and the public repertoire

Having considered the foundations of T cell immunity in terms of antigen presentation, TCR diversity and MHC-peptide-TCR binding, it is important to address some of the broader systemic properties of the T cell response. There are many complexities, and most of these lie outside the scope of this thesis, where the emphasis is firmly on antigen presentation and recognition. One such area relates to T cell differentiation. Three major subpopulations of T cell have been mentioned in earlier sections — $CD8^+$ T cells, $CD4^+$ T cells and regulatory T cells — but these can be further subdivided into distinct subsets with different functions. Taking just one example, $CD4^+$ (or helper) T cells can be divided into two major subsets — T_H1 cells that may be regarded as "classical" $CD4^+$ T cell, and T_H2 cells that typically target extracellular pathogens such as parasitic worms (helminths) (Walker & McKenzie, 2018) – but several other $CD4^+$ T cell subsets have been identified. (A list of distinct subsets of T cells that are currently recognised and the complex signalling events that trigger their emergence are discussed in a recent review [Broere & van Eden, 2019].)

Nevertheless, there are at least two systemic features of T cell responses that are highly relevant to this research. The first concerns concerned the breadth of the response to a given challenge: there may be many peptide-MHC combinations that an individual's TCRs are capable of binding to, but the response is typically focused on a small number of peptide-MHC complexes, and perhaps only one. This characteristic is known as *immunodominance* and has been observed in both $CD8^+$ T cells (Yewdell, 2006) and $CD4^+$ T cells (Sant et al., 2007). Although many of the factors contributing to immundominance have been identified, the balance of factors that determine which TCR-peptide-MHC complexes emerge as dominant in a particular context is far from clear, and subdominant responses may also afford protection from disease (Tscharke et

al., 2015). One factor known to make a partial contribution – and one that is available (via prediction) in this research (see section 2.2) – is the affinity with which the epitope binds to the MHC molecule (Kotturi et al., 2008).

It has long been recognised that individuals sharing the same HLA allele often target the same antigenic peptide, with potential implications for vaccine design (Chen & McCluskey, 2006; Sant et al., 2007). For example, individuals having a human cytomegalovirus (HCMV) infection that share the very common MHC class I allele HLA-A*02:01 have $CD8^+$ T cell responses that mainly target a single HCMV peptide from protein pp65 (Wills et al., 1996). In some cases, such as HCMV, the T cells that target these immunodominant peptides have public TCR sequences (i.e., sequences that are shared between multiple individuals). (For an analysis of the public TCRs that target the HCMV pp65 immunodominant epitope, see [Yang et al., 2015].) Underpinning the occurrence of public TCRs targeting a specific antigen is the high frequency of public TCRs within the naïve T cell repertoire, which (notwithstanding the high levels of TCR diversity) is attributable to biases in the V(D)J recombination process (discussed in section 1.2.2) and to the impact of central tolerance mechanisms (discussed in section 1.2.1) (Shugay et al., 2013).

1.3 Overview of research

The aims of the research presented in this thesis is to develop and refine computational strategies for applying existing MHC binding prediction tools to address biomedical problems that share two key characteristics: firstly, they arise close to the boundary between self and non-self; and secondly patient outcomes are stratified with respect to their personal complement of HLA alleles. In the simplest case, a single residue difference between a self-protein and a non-self "alternative" protein may or may not be detectable by the host immune system depending on that host's specific set of HLA alleles, determining whether or not that host has the potential to mount a mature immune response to that "alternative" protein. Three contrasting examples of "alternative" protein are explored in this research: a replacement therapeutic (Factor VIII in chapter 3); mismatched MHC molecules in organ transplantation (chapter 4); and tumour antigens (hepatocellular carcinoma in chapter 5).

To conclude this chapter (before moving on to consider the specific computational methods developed for this research and their areas of application), it is worth considering the underlying "philosophy" of this research, bearing in mind the various background topics discussed in preceding sections.

The approaches developed for this research are consistent in three important respects. Firstly, they are applied to tasks where the efficacy of experimental methods on their own is limited by the scale of the combinatorial challenge that the tasks entail. For example, the first application – missense mutation haemophilia A (chapter 3) – entailed the evaluation of over 4 million peptide-MHC combinations.

Secondly, the focus is on aspects of the T cell response where computational methods are sufficiently accurate. Given the current limitations of TCR binding prediction (discussed in section 2.3), the focus here was on the prediction of peptide-MHC binding, but the extent to which peptide-MHC binding can be deemed "sufficiently accurate" depends on the context. Taking as an example the third application – the selection of potential diagnostic biomarkers for hepatocellular carcinoma from a large pool of antigenic peptide candidates (chapter 5) – the minimum requirement was that peptide-MHC binding prediction is sufficiently better than random to justify the (non-trivial) time taken and (very limited) computational resources required.

Thirdly, in handling areas of uncertainty about T cell responses, this research has generally made simplifying, and rather conservative, assumptions. For example, although tolerance mechanisms sometimes fail – and, given the high incidence of certain autoimmune diseases, this is clearly not a rare occurrence – in the absence of an ability to predict such occurrences, the practical assumption for the antigens of interest is that tolerance always works. There are likely to be additional holes in an individual's T cell repertoire (i.e., presented peptides for which no TCRs capable of binding to the corresponding peptide-MHC surface exist), but in the absence of an ability to predict where such holes may occur, the working assumption is that only central tolerance creates holes in the coverage of TCRs. And although a given TCR will not necessarily make contact with all the TCR-facing residues of an antigenic peptide, in the absence of an ability to predict which of these residues (if any) are not in contact with the TCR, the working assumption is that a TCR makes contact with all of the TCR-facing residues.

2 Computational Methods and Resources

2.1 Data resources

The methods used and developed for the research presented here are computational, but there is an underlying reliance on experimental data during the development process and for the evaluation of performance. A variety of different kinds of data relevant to this research is available in a set of key public resources.

Vital information about human MHC molecules is available from two key sources. Firstly, the official sequences of human MHC molecules named by the appropriate WHO nomenclature committee are available from the IPD-IMGT/HLA database (https://www.ebi.ac.uk/ipd/imgt/hla/) (Robinson et al., 2020). As already noted in section 1.1.4, as of 24th April 2022 the IPD-IMGT/HLA database contains more than 24,000 HLA class I alleles and 9,000 HLA class II alleles. Secondly, the frequencies with which HLA alleles occur in different populations is stored in the Allele Frequency Net Database (http://allelefrequencies.net), with data collected from various sources, including peer-reviewed publications, dedicated workshops and individual lab submissions (Gonzalez-Galarza et al., 2020). Data can be interrogated at various levels of granularity, ranging from individual studies (which may focus on a particular region or minority within a single country) to worldwide aggregations with respect to a geographical region or ethnic group. As of 24th April 2022, the database has HLA information collated from nearly 1,300 population studies, with data collected from over 14 million individuals.

The largest resource for information about experimentally verified epitopes, including T cell epitopes, is the Immune Epitope Database (IEDB, https://www.iedb.org/), with information mainly collected from peer-reviewed papers and "manually curated followed structured curation guidelines" (Vita et al., 2019) The available metadata varies considerably between epitopes, but commonly includes the source organism and location of the epitope within an antigenic protein (identified by name and UniProt ID), the associated host species and MHC alleles, and the source citations and type of experimental assays undertaken. As of 24th April 2022, the IEDB contains nearly 640,000 HLA class I entries and over 450,000 HLA class II entries.

A second resource containing epitope sequences is SYFPEITHI (http://www.syfpeithi.de/) (Rammensee et al., 1999). Although it contains far fewer epitope sequences than the IEDB, it provides a useful MHC-oriented perspective, deriving information about MHC-specific motifs — that is, residue preferences at anchor (binding pocket) and other positions within the MHC groove. Whereas the SYFPEITHI motifs specify amino-acid types belonging to specific categories (e.g., anchor residues), the MHC Motif Viewer (https://services.healthtech.dtu.dk/services/MHCMotifViewer/Home.html) provides MHC-specific sequence logos and position-specific scoring matrices (Rapin et al., 2008).

Information about TCRs is available from various sources. Information about the germline genes that encode TCRs is available from the IMGT reference directory (https://www.imgt.org/vquest/refseqh.html) (Lefranc & Lefranc, 2001). TCR sequences from individual T cell repertoires are available from various repositories, many of which can be queried in an integrated manner via the iReceptor Gateway (http://ireceptor.irmacs.sfu.ca/) (Corrie et al., 2018). A typical T cell repertoire dataset contains tens of thousands of sequences — often just β chain sequences — each having its CDRs and germline genes annotated. Although many such repertoires are associated with specific diseases, there is no information about the antigenic targets of individual sequences. The largest source of information about known ternary MHC-peptide-TCR complexes is VDJdb (https://vdjdb.cdr3.net/). VDJdb is a manually curated database in which each TCR entry contains information about the TCR α and/or β chain V and J genes together with their CDR3 sequences, the associated MHC allele, the sequence of the bound epitope, and other relevant metadata (Bagaev et al., 2019). As of 24th April 2022, VDJdb contains entries for nearly 40,000 human TCR β chains, of which over 20,000 have information about their corresponding α chains. Note, however, that the number of unique epitopes is much smaller — only 1,087 for the complete set of TCR β chains.

The primary database for solved 3D structures of biomolecules is the Protein Data Bank (PDB, http://rcsb.org) (Berman et al., 2000). IMGT/3Dstructure-DB (https://www.imgt.org/3Dstructure-DB/) is a secondary database that incorporates PDB-derived structures of MHC molecules, peptide-MHC complexes, and the variable regions of TCRs either unbound or forming a ternary complex with peptide and MHC

(Ehrenmann et al., 2010). As of 24th April 2022, IMGT/3Dstructure-DB contains over 950 human MHC structures (of which more than 720 are in complex with peptides) and over 350 TCR structures (of which 180 are part of a ternary complex with peptide and MHC).

2.2 Antigen presentation prediction methods

Of the various aspects of antigen presentation and the engagement between MHC molecules, peptides and TCRs, by far the most widely addressed challenge using computational methods is that of predicting whether a given peptide binds to a given MHC molecule with sufficient affinity to be a candidate T cell epitope (subject to the presence of a TCR capable of binding that peptide-MHC complex) — a task that is generally referred to as T cell epitope prediction or MHC binding prediction. Given the combinatorial challenges associated with large numbers of both HLA alleles and novel antigens, and given that MHC binding experiments are both costly and time-consuming, the attractions of using computational prediction methods are clear.

MHC binding predictors are typically developed using information about known epitopes from curated databases such as the IEDB and/or SYFPEITHI (see section 2.1). It is notable that peptides with a negative outcome (i.e., peptides experimentally determined not to bind to a given MHC molecule) are under-reported in the literature and under-represented in curated databases. Hence, whereas the IEDB contains (as of 24th April 2022) over 1 million positive peptide-MHC entries, there are less than 100,000 negative entries. In reality, only a small fraction of peptides are binders, although the numbers vary between MHC molecules. Consequently, method developers commonly make the assumption that, if a set of binding peptides have been identified for a given combination of MHC molecule and antigen, all other peptides from that antigen are non-binders with respect to the same MHC molecule.

As noted in a recent review, several motif- and matrix-based prediction methods were developed in the 1990s, but more accurate predictions because possible from the late 1990s with the creation of epitope databases (notable, in 2003, the IEDB, as described in section 2.1) that contained sufficient data for training machine learning methods — notably artificial neural networks (ANNs) (Peters et al., 2020). Another key advance, given the highly polymorphic nature of the HLA locus, was the development of pan-

specific methods capable of making predictions for MHC molecules for which there are insufficient known epitopes to support conventional training. Such methods utilise information about the amino acids that occur within the binding groove and/or pockets of the MHC molecule (Peters et al., 2020).

A 2015 survey listed 20 different T cell epitope predictors (Soria-Guerra et al., 2015). The fair evaluation of such methods is challenging, as it requires data that was not used in the training of any of the methods to be available for testing. Within the past decade, this challenge has been overcome for an important subset of methods that are available as online servers via automated benchmarking frameworks for MHC class I (Trolle et al., 2015) and class II (Andreatta et al., 2018). The benchmark is run weekly using new data in the IEDB prior to its public release, with results made available on the IEDB website (http://tools.iedb.org/auto_bench/mhci/weekly/ and http://tools.iedb.org/auto_bench/mhcii/weekly/ respectively). In both cases, a threshold for positive predictions is set as < 500 nmol/L for IC_{50} binding measurements and over 2 hours for the half-life of binding measurements.

Prediction tools developed at the Technical University of Denmark (DTU) — notably NetMHC (Andreatta & Nielsen, 2016; Nielsen et al., 2003), NetMHCpan (Hoof et al., 2009), NetMHCII (Jensen et al., 2018) and NetMHCIIpan (Reynisson et al., 2020) — have consistently performed well in benchmark assessment for many years, are widely used, and have been the tools of choice in this research. The pan-specific versions of the tools have been evaluated by the IEDB automated benchmarking framework since its inception and have accumulated AUC scores of over 0.85 for most HLA class I alleles and over 0.8 for a set of key HLA class II alleles[1]. HLA class II prediction is generally considered more difficult than class I prediction because it is necessary to predict the register of peptide binding within the open groove of the MHC class II molecule and needs to take into account the effect of the flanking regions of the peptide that lie outside the groove. It has recently been suggested that "MHC class II binding predictions have broadly caught up to where MHC class I binding predictions were a decade ago" (Peters et al., 2020).

[1] Note that predictions for certain HLA class II alleles are absent from the benchmark report for reasons that are unclear.

Additional aspects of the MHC class I antigen presentation pathway that are amenable to computational prediction are TAP transport efficiency and C-terminal proteasomal cleavage. Methods for predicting both have been integrated into two additional DTU tools, NetCTL (Larsen et al., 2007) and NetCTLpan (Stranzl et al., 2010), with cleavage prediction also available as a stand-alone tool, NetChop (Nielsen et al., 2005; Kesmir et al., 2002). In both tools, TAP transport efficiency and C-terminal proteasomal cleavage are given low weighting in terms of their contribution to the overall score (0.05 for TAP transport efficiency and 0.15 for C-terminal cleavage in the case of NetCTL).

2.3 TCR binding prediction

TCR binding prediction is intrinsically more challenging than MHC binding prediction (section 2.2). In the latter, a peptide adopts an extended conformation within an essentially static MHC binding groove (see section 1.1.3), whereas TCRs adopt somewhat different orientations to peptide-MHC surfaces and binding involves the engagement of up to six flexible loops – the CDRs (see section 1.2.3). Moreover, the number of known TCR-peptide-MHC complexes is much smaller than the number of known peptides-MHC complexes (see section 2.1), making it more difficult to understand the "rules" of TCR binding (in so far as they exist) or train machine learning algorithms to make predictions.

Given the extent of the challenge and the limited amount of data – most particularly paired TCR chain data – the task of predicting TCR binding from sequence has often been formulated in somewhat restrictive ways. Generally, the task is treated as a binary, binding/non-binding problem, with no attempt made to predict binding affinity. Certain methods only make predictions for specific peptides, bearing in mind that many TCRs can bind to the same epitope, and some utilise only the TCR β chain that typically makes dominant contact with the peptide (as noted in section 1.2.3). As an example, all the preceding restrictions apply to the web tool TCRex (https://tcrex.biodatamining.be) (Gielis et al., 2019).

One notable contribution to this space has been the development of NetTCR (the latest version, 2.0, is available at https://services.healthtech.dtu.dk/service.php?NetTCR-2.0) by the same research group that developed the NetMHC tools (Montemurro et al., 2021). NetTCR uses a type of deep learning architecture known as a convolutional neural

network (CNN). The authors experimented with different input data and concluded that "TCR-peptide interactions can only to a very limited extent be characterized using current CDR3β peptide data" (Montemurro et al., 2021). The degree of similarity between data used for training and evaluation makes the fair measurement of predictive performance particularly challenging. Using a NetTCR model that utilises paired α and β chain data, performance largely depended on the number of TCRs associated with a given peptide: the AUC for peptides characterised by "200 or more" TCRs was 0.88, but only 0.38 for those characterised by "20 or fewer" TCRs (Montemurro et al., 2021). Given the kinds of scenario addressed by the research presented in this thesis, where there is no expectation that a given epitope will have any TCRs that are known to bind to it, these conclusions appear distinctly unpromising. TCR binding prediction was not undertaken for this research.

If a breakthrough in predictive performance occurs in the near future, it appears likely to involve the exploitation of advanced deep learning approaches that have proved highly effective in the field of Natural Language Processing (NLP) and are being used increasingly in address challenges in other areas of biology, with AlphaFold 2 as the most notable example (Jumper et al., 2021). One such NLP-derived strategy is called BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), which has been adapted to address the TCR binding problem in the recently released TCR-BERT (Wu et al., 2021) and TCRBert (Han & Lee, 2021). Undertaking a fair benchmark assessment of these new methods is a current research goal for Prof. Adrian Shepherd's research group.

2.4 Repository

The source code for the computational pipeline to undertake the analysis and post processing of the results for the various applications presented in chapter 3, chapter 4 and chapter 5 can be viewed at the GitHub repository https://github.com/nuzun/NetPredictionApplication.

3 Predicting Inhibitor Risk in Missense Mutation Haemophilia A

3.1 Introduction

Haemophilia A (HA) is an X-linked hereditary disorder, which results from deficiency of Factor VIII (FVIII), one of the blood clotting proteins. The X chromosome involves the genes coding for FVIII, hence HA is overwhelmingly a condition that affects men (approx. 1 in 5000 male births), whereas women are carriers. The activity level of the FVIII determines the severity of the disease: severe when < 1%, moderate when 1-5%, and mild when > 5%. The standard treatment for patients diagnosed with HA is replacement FVIII – either recombinant or plasma derived – to prevent or treat bleeding episodes. HA patients mostly have bleeding in soft tissue, the joints or muscles. Whereas severe HA patients may have frequent and spontaneous bleeds from infancy, bleeding in non-severe patients is typically associated with injury or medical procedure such that first treatment with therapeutic FVIII (tFVIII) may be delayed into, or beyond, middle age.

For patients with HA of all severities, by far the most important risk to the effectiveness of tFVIII treatment is the development of anti-FVIII antibodies, known as inhibitors. Inhibitor formation occurs in 25-30% of severe HA patients and 7-15% in non-severe group (Lieuw, 2017). Such patients typically require alternative treatment regimens involving costly FVIII immune tolerance induction (ITI) to suppress the immune response to FVIII, or more recently, the use of bypassing agents such as activated recombinant Factor VII to control bleeding (Konkle et al., 2007). In around half of non-severe patients, inhibitors cross-react with the patient's endogenous FVIII result in a more severe disease phenotype (Hay et al., 1998).

For all HA patients, the ability to predict the risk of inhibitor formation has potential therapeutic advantages. For example, severe HA patients with high inhibitor risk might be candidates for pre-emptive ITI, whereas non-severe HA patients might be offered alternative therapeutics, such as desmopressin (which stimulates the release of von Willebrand factor, which in turn reduces the rate at degradation rate of FVIII in the bloodstream) (Fanchini & Mannucci, 2011). Indeed, accurate prediction of inhibitor risk in non-severe patients would be of immediate therapeutic advantage (Dr Dan Hart, Barts, personal communication).

In this research, the primary focus is on HA patients with a single missense mutation. Missense mutation HA is associated with disease phenotypes of all severities, depending on the location of a given mutation within the FVIII molecule, but most commonly with non-severe HA. As noted in the paper to which this research contributed, "inhibitor screening in the setting of non-severe hemophilia A is currently more reactive and sporadic [than for severe HA] but recognized to be of increasing importance given the aging population of those living with non-severe hemophilia A" (Hart et al., 2019).

Inhibitor formation is a $CD4^+$ T cell dependent process (Jacquemin et al., 2003). The missense mutation in the patient's endogenous FVIII is the trigger for the patient's immune system in recognizing the tFVIII-derived peptides as foreign, notably those peptides spanning the location of the missense mutation. Peptides bound to MHC class II molecules are presented to $CD4^+$ T cells by professional antigen presenting cells (as described in section 1.1). This may lead to $CD4^+$ T cell activation and ultimately the formation of inhibitors; however, the immune response towards tFVIII-derived peptides that are undistinguishable from the patient's endogenous peptides is supressed by self-tolerance mechanisms; in particular, self-reactive $CD4^+$ T cells are removed from the naïve repertoire in the thymus by central tolerance mechanisms (discussed in section 1.2.1).

Three key factors determine whether a given tFVIII-derived peptide triggers a T cell response. Firstly, the peptide may, or may not, be presented, depending on whether it is capable of binding to any of the individual's MHC molecules (as encoded by his HLA class II genes) with sufficient avidity (as discussed in section 1.1.3). Secondly, if presented, the peptide may, or may not, form a novel peptide-MHC surface, i.e., one that is distinguishable from those formed by self-peptides; only if such a surface differs from those formed by endogenous peptides is a $CD4^+$ T cell response normally possible. Thirdly, if a novel peptide-MHC surface is presented, the individual's $CD4^+$ T cell repertoire may, or may not, contain T cells capable of binding to that surface – although their presence is by no means improbable given the diversity of T cell receptors within the repertoire (see section 1.2.2) and TCR cross-reactivity (discussed in section 1.2.3). Whereas the first two factors are potentially testable in advance, the third factor depends on the (at least partially) stochastic V(D)J recombination process by which T cell receptors are generated and may vary over time. Hence prediction of (potential) risk depends on predicting the formation, or otherwise, of novel peptide-MHC surfaces.

Given the highly polymorphic nature of HLA genes (discussed in section 1.1.4) together with the hundreds of FVIII missense mutations that have been reported so far, and given the relatively low-throughput characteristics of peptide-MHC binding assays, undertaking a large-scale *in vitro* analysis of inhibitor risk is infeasible. The aim of this research is to address this challenge utilising *in silico* techniques, building on an approach developed previously in Dr Shepherd's research group at Birkbeck, University of London (Shepherd et al., 2015). The key innovation of the work presented here was to greatly extend the search for potential cross-matches between non-self (a given therapeutic peptide) and self; originally only the corresponding FVIII location was considered, whereas in this research cross-matches to other locations in endogenous FVIII and to other proteins in the human proteome were considered. Preliminary analyses were undertaken by several individuals, but all the final code and results published in (Hart et al., 2019) were generated by me.

3.2 Methods

3.2.1 The identification of novel peptide-MHC surfaces

The underlying methodology for predicting inhibitor development risk of patients with missense mutation HA has been well described in (Shepherd et al., 2015). Briefly, all tFVIII (UniProt FVIII sequence P00451(The UniProt Consortium, 2016)) 15-mers that span the location of the missense mutation were analysed NetMHCII, one of the accurate MHC binding predictors developed by the Technical University of Denmark (DTU) that are listed in section 2.2. For this research, a downloadable version of NetMHCII 2.2 was used (Nielsen and Lund 2009); this has now been superseded by NetMHCII version 2.3 (available at https://services.healthtech.dtu.dk/service.php?NetMHCII-2.3). Any such peptide that was predicted to i) have a binding register that places the location of the missense mutation within the 9-residue binding groove of the MHC molecule and ii) bind to the MHC molecule with sufficient affinity was considered a candidate for forming a novel peptide-MHC surface. In this research, the "sufficient affinity" threshold was considered to be IC_{50} < 1000 nmol/L – as noted in section 1.1.3, a widely accepted threshold taken to indicate biologically relevant binding (Southwood et al., 1998; Paul et al., 2015; Paul et al., 2020). A peptide was subsequently predicted to form a novel peptide-MHC surface if either:

a) The corresponding endogenous peptide (i.e., with the equivalent predicted binding core) was a non-binder. This situation may arise when the missense mutation is at an MHC facing position; for most HLA alleles, these are positions 1, 4, 6 and 9 within the binding groove. In this case, it was predicted that T cells capable of binding to the peptide-MHC complex will not have been eliminated by self-tolerance mechanisms; or

b) The residue difference between endogenous FVIII and tFVIII was at a TCR-facing position, namely at position 2, 3, 5, 7 or 8 within the binding groove. Hence the presented surface is different to that presented by the same MHC molecule bound to the endogenous peptide. These scenarios are summarised in **Figure 3.1**A.

For this research, the original set of 14 HLA-DR alleles (Shepherd et al., 2015) was expanded to incorporate common HLA-DP and HLA–DQ alleles, giving a total of 25 HLA class II alleles. This set has been used in previous MHC class II research because of their estimated global population coverage: greater than 70% for HLA-DR alleles, greater than 90% for HLA-DP alleles, and greater than 80% for HLA-DQ alleles (Wang et al., 2010).

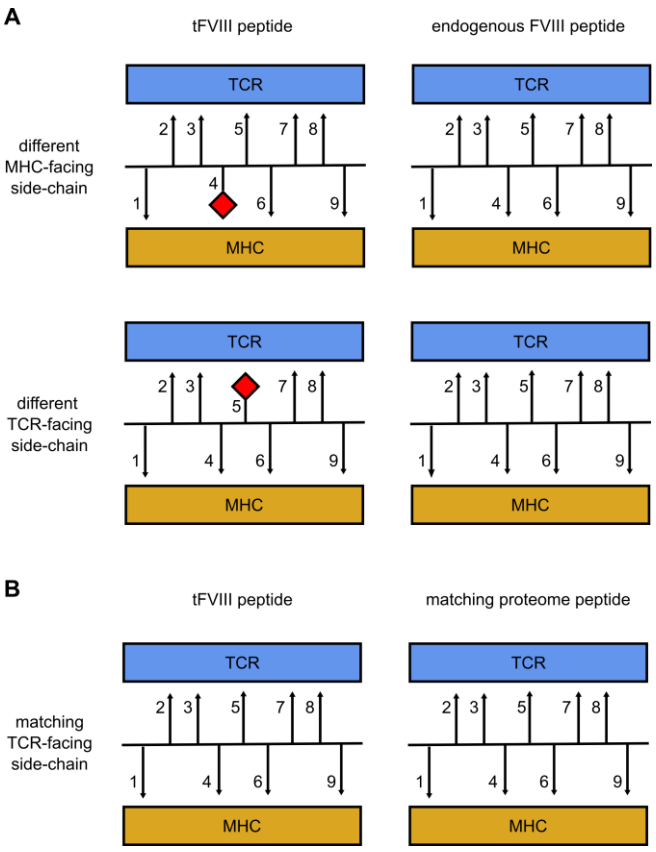


Figure 3.1 Schematic diagram explaining how side-chain differences can lead to the presentation of novel peptide-MHC surfaces. The FVIII missense mutation is denoted by a red diamond (figure adapted from [Hart et al., 2019] with permission to reuse the material granted by Ferrata Storti Foundation). A) Image depicts the scenarios how a FVIII missense mutation can be at different positions along the peptide affecting the prediction of forming a novel-peptide-MHC surface. B) Following the preceding assessment, if a tFVIII peptide is associated with the potential formation of a novel peptide-MHC surface, such a surface will not be novel if there is a peptide from a different location within the human proteome that is a binder and has the same TCR-facing residues.

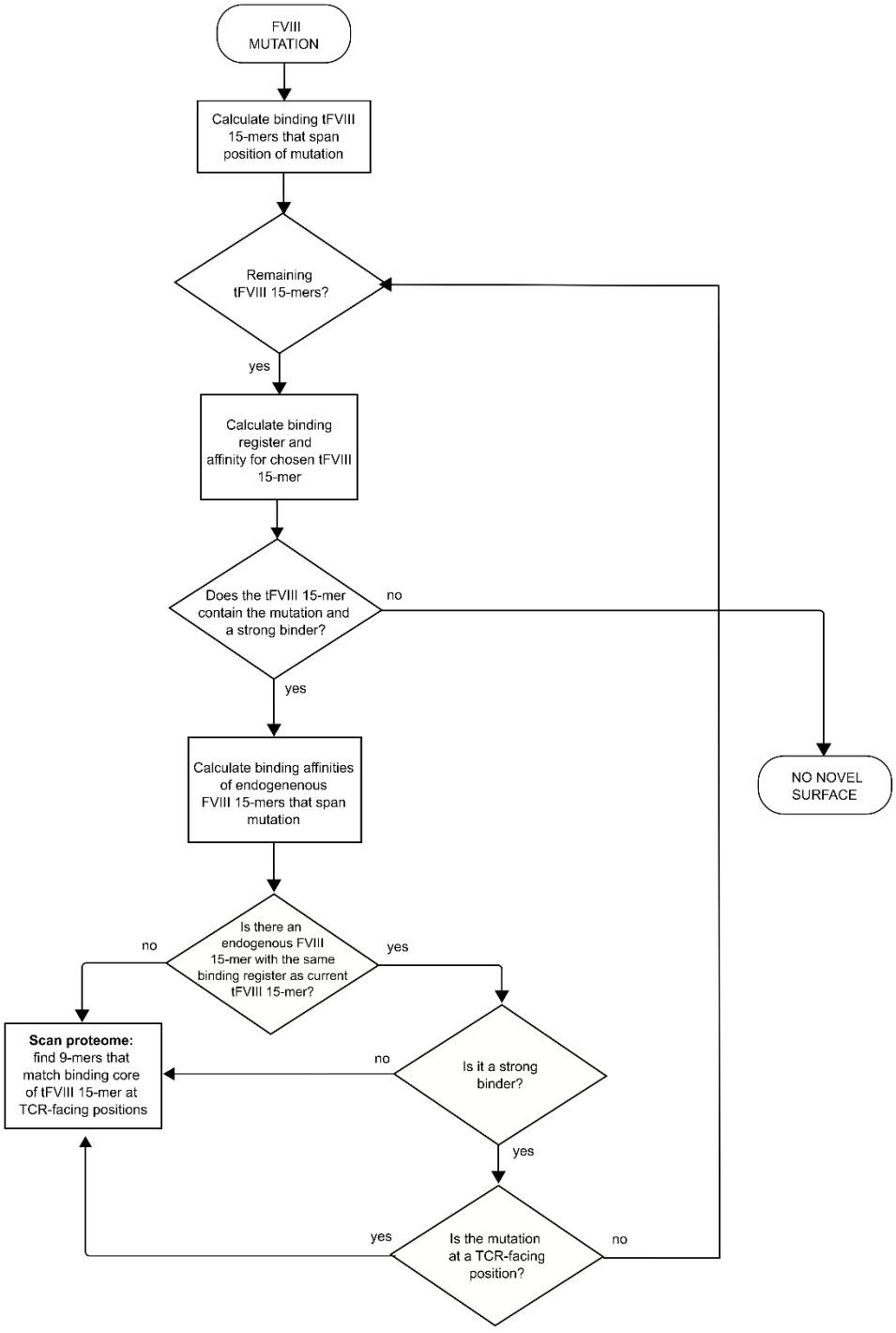
3.2.2 Scanning novel peptides against the human proteome

The main innovation of the research presented here involved taking into account the possibility that some of the apparently novel peptide-MHC surfaces identified by the preceding approach may not be associated with an inhibitor risk because of cross-matches to self-peptides other than those spanning the equivalent location in the individual's endogenous FVIII. This scenario is summarised in **Figure 3.1**B. To undertake this analysis, the core 9-mer from the peptide forming a putative novel peptide-MHC complex was scanned against more than 11 million unique 9-mers forming the human proteome, derived from the more than 100,000 human protein sequences (including isoforms) retrieved from Ensembl database (Yates et al., 2016). Each relevant cross-match

(i.e., at positions 2, 3, 5, 7 and 8 of the 9-mer) was then analysed; if any 15-mer spanning a cross-matching 9-mer was predicted to bind with i) that 9-mer as its binding core and ii) with sufficient affinity, the original peptide-MHC surface was no longer deemed to be novel, and hence no longer associated with inhibitor risk.

This approach was termed *proteome scanning* and the peptides with eliminated risk were considered *proteome protected*. In cases where multiple novel peptide-MHC surfaces were identified for a single missense mutation and only a subset was associated with proteome protection, that mutation was considered *partially proteome protected*. **Figure 3.2** summarizes the computational pipeline used in this analysis.

A


FVIII MUTATION
Calculate binding tFVIII 15-mers that span position of mutation
Remaining tFVIII 15-mers?
yes
Calculate binding register and affinity for chosen tFVIII 15-mer
Does the tFVIII 15-mer contain the mutation and a strong binder?
no
yes
NO NOVEL SURFACE
Calculate binding affinities of endogenous FVIII 15-mers that span mutation
Is there an endogenous FVIII 15-mer with the same binding register as current tFVIII 15-mer?
no
yes
Scan proteome: find 9-mers that match binding core of tFVIII 15-mer at TCR-facing positions
Is it a strong binder?
no
yes
Is the mutation at a TCR-facing position?
yes
no

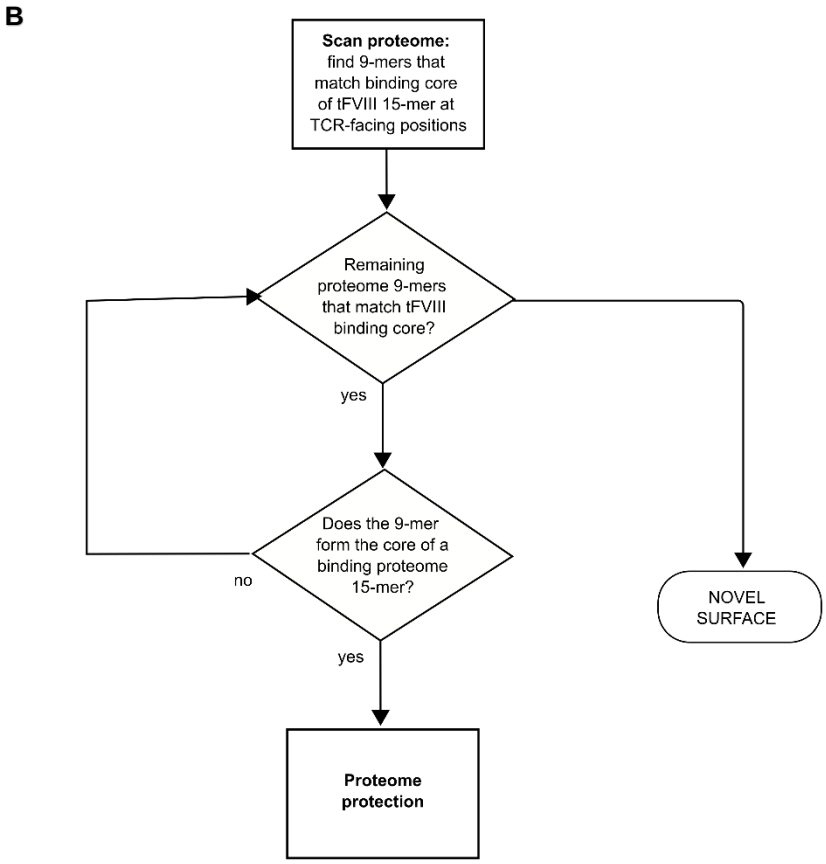


Figure 3.2 Flowcharts showing how the assessment of HA inhibitor risk is undertaken for a given FVIII missense mutation, taking into account potential cross-matches to the human proteome where necessary (A) and elaboration of the proteome scanning step (B) (figures adapted from [Hart et al., 2019] with permission to reuse the material granted by Ferrata Storti Foundation). Following the standard usage of flowchart symbols, "pill" or "stadium" shapes represent start or end points, rectangles represent processes, diamonds represent decisions (yes or no), and arrows represent the direction of flow.

3.2.3 Evaluating statistical significance

Patient data was retrieved from the Factor VIII Gene Variant Database of European Association for Haemophilia and Allied Disorders, EAHAD (http://www.factorviii-db.org, accessed on November 26, 2016). Data was filtered based on the patients' inhibitor formation status and excluded patients with an inhibitor formation status "unknown". There were 2,225 individuals with 956 distinct FVIII mutations reported at 605 different locations. The number of patients with inhibitors was 160.

The two-tailed Fisher's Exact Test was implemented in the R statistical programming language to evaluate the strength of the predictions against the real patient data with a null hypothesis that predicted inhibitor formation rates and rates calculated from the patient data are independent at the 0.05 p-value of significance. In this research, a revised method to determine a patient's predicted inhibitor formation risk was introduced. In previous research, an "unknown risk" category was not defined for evaluation purposes, whereas for this research a missense mutation is assigned to one of three predicted-degree-of-risk categories: "low/negligible risk", "at risk" or "unknown risk". A missense mutation is considered in the "at risk" category if risk is predicted for one or more of the 4 HLA class II allele sets: HLA-DRB1 (11 alleles), HLA-DRB3/4/5 (3 alleles), HLA-DP (5 alleles) and HLA-DQ (6 alleles). Each gene set is predicted to be associated with risk if no more than a single allele in the set is predicted to be no-risk (here we assume that all individuals are heterozygous). A patient is predicted to be at "low/negligible" risk of inhibitor formation if all HLA allele/mutation combinations are predicted to be associated with no risk. The risk status of any other patient is considered "unknown"; all such patients are excluded from the statistical calculations.

3.3 Results

3.3.1 A proteome scanning example

Here, a single example of proteome scanning is worked through in detail. The example combines the FVIII missense mutation Arg593Cys - R593C is the "traditional" numbering based on the mature protein that appears in most publications. The equivalent numbering used by the Human Genome Variation Society (https://www.hgvs.org/) is R612C - with the common HLA-DR allele HLA-DRB1*01:01. R593C was chosen because it is a relatively common FVIII missense mutation that is reportedly associated with a comparatively high risk of inhibitor formation. For example, one major study containing 106 individuals with the R593C mutation reported that 12 of these individuals (or 11.3%) developed inhibitors (Eckhardt et al., 2013).

Step one involves predicting which, if any, tFVIII 15-mers that span position 593 are predicted to be binders using NetMHCII (Nielsen & Lund, 2009). In this case, two binding cores (i.e., 9-mers positioned within the MHC groove) that span position 593 –

IQRFLPNPA and YLTENIQRF – were associated with multiple predicted binding 15-mers, as shown in **Figure 3.3**A.

The second step involves assessing whether any cores are predicted to form a novel peptide-MHC surface in comparison with the surfaces formed by their respective endogenous peptides. In this case, both preceding endogenous cores – with a Cys (C) replacing the Arg (R) at positions 3 and 8 respectively – belong to predicted binding 15-mers. Given that both of these positions are TCR facing (bearing in mind that the MHC binding pockets associated with the HLA-DRB1*01:01 allele are at positions 1, 4, 6 and 9), both cores are deemed to form novel peptide-MHC surfaces (as shown in **Figure 3.3**B) and hence – at this stage of the analysis – constitute a risk in terms of potential inhibitor development.

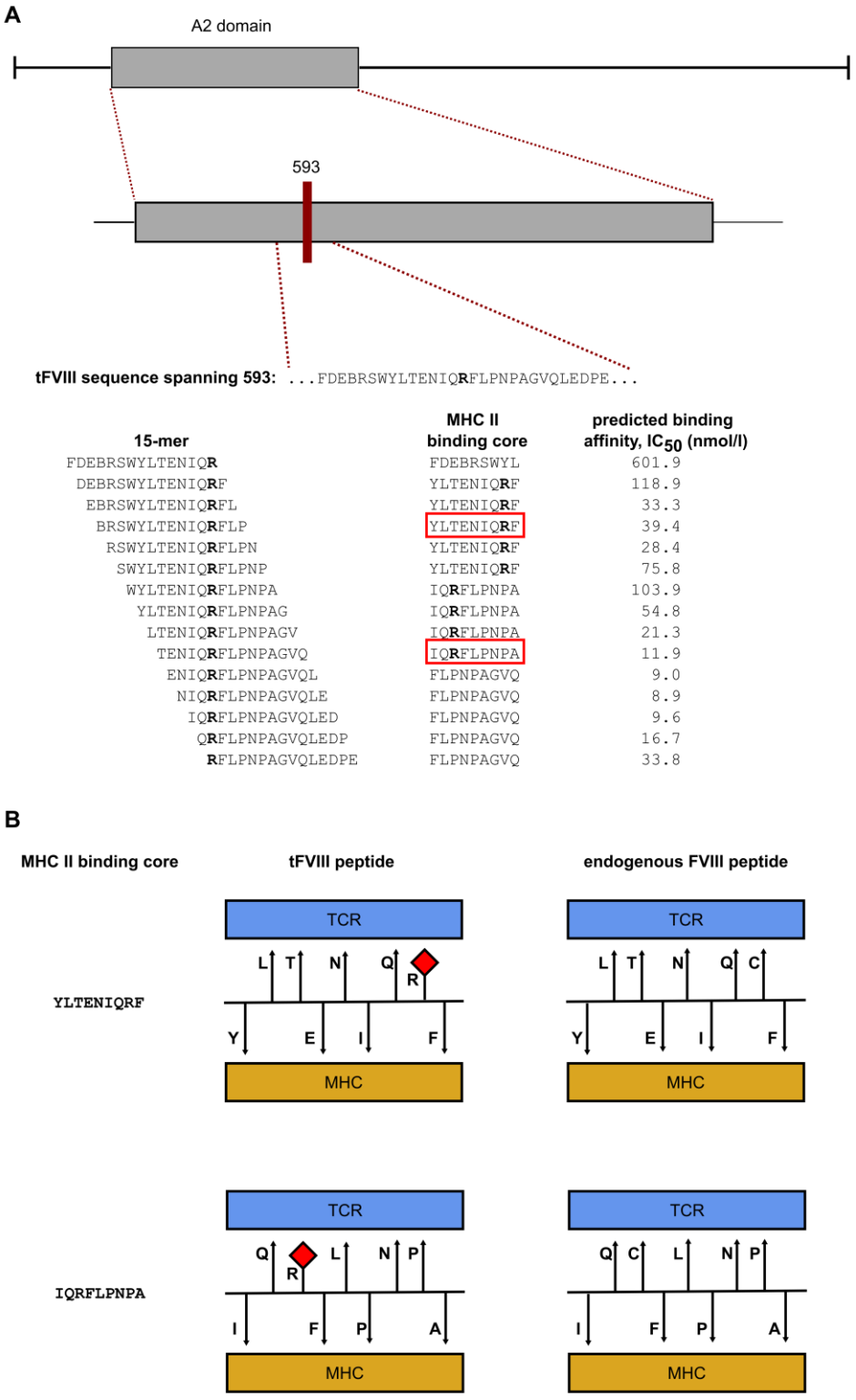


Figure 3.3 An example of novel peptide-MHC surface formation, using the combination of Arg593Cys and allele HLA-DRB1*01:01 as an example (figure adapted from [Hart et al., 2019] with permission to reuse the material granted by Ferrata Storti Foundation). A) 15-mers from tFVIII containing two cores that span Arg593 (R593) – IQRFLPNPA and YLTENIQRF – are predicted to bind to the MHC molecule associated with allele HLA-DRB1*01:01 by NetMHCII. B) As both cores have R593 at a TCR-facing

position, both are predicted to form peptide-MHC surfaces that are novel in comparison to those formed by the patients' endogenous FVIII, which have a TCR-facing Cys593 (C593).

The third and fourth steps involve proteome scanning and are designed to establish whether the peptide-MHC surfaces that (by the end of step two) are deemed novel with respect to the location of the missense mutation within the individual's endogenous FVIII are also novel in the wider context of their other proteins. Given that matching peptides is much quicker that predicting MHC binding, step three involves hunting for proteome 9-mers that match the binding cores of interest at their TCR-facing positions. In the current example, this corresponds to scanning against the pre-calculated library of 11,272,502 unique proteome 9-mers using the patterns XQRXLXNPX (for core IQRFLPNPA) and XLTXNXQRX (for core YLTENIQRF), where each X is at an MHC-facing position and matches any amino-acid type. In this case, pattern XQRXLXNPX matches the 9-mer FQRELNNPL found in human tubulin polyglutamylase (UniProt sequence Q6ZT98), and pattern XLTXNXQRX matches the 9-mers GLTENSQRD and ELTKNAQRA found in dystrobrevin binding protein 1 (dysbindin) (UniProt sequence D6RJC6) and uncharacterized human protein C2orf48 (UniProt sequence Q96LS8) respectively (as shown in **Figure 3.4**A). Tubulin polyglutamylase is an enzyme which is highly expressed in the nervous system including the spinal cord, thalamus, hippocampus, hypothalamus, and cerebellum (Ikegami et al., 2006). Similarly, dysbindin is expressed in the prefrontal cortex and hippocampus and has been identified as one of the susceptibility genes for schizophrenia. It modulates prefrontal brain functions and is involved in neuronal development (H. Wang et al., 2017). Given their critical and life-long functional roles, there is no reason to expect that either protein is likely to "avoid" thymic central tolerance.

A

FVIII:	IQRFLPNPA
matching pattern:	XQRXLXNPX
tubulin polyglutamylase:	FQRELNNPL
FVIII:	YLTENIQRF
matching pattern:	XLTXNXQRX
dysbindin:	GLTENSQRD
C2orf48:	ELTKNAQRA

B

15-mer	MHC II binding core	predicted binding affinity, IC_{50} (nmol/l)
tubulin polyglutamylase matches to pattern XQRXLXNPX:		
SGRAASFQRELNNPL	FQRELNNPL	26.3
GRAASFQRELNNPLK	FQRELNNPL	13.8
RAASFQRELNNPLKR	FQRELNNPL	7.9
AASFQRELNNPLKRM	FQRELNNPL	5.3
ASFQRELNNPLKRMK	FQRELNNPL	6.4
SFQRELNNPLKRMKE	FQRELNNPL	8.2
FQRELNNPLKRMKEE	FQRELNNPL	11.3
dysbindin matches to pattern XLTXNXQRX:		
MSSPGLTENSQRDPS	GLTENSQRD	8207.8
SSPGLTENSQRDPSE	GLTENSQRD	7036.4

Figure 3.4 An example of proteome cross-matching, using the combination of Arg593Cys and allele HLA-DRB1*01:01 as an example (figure adapted from [Hart et al., 2019] with permission to reuse the material granted by Ferrata Storti Foundation).

The fourth and final step involves checking whether any of the proteome cross-matching 9-mers occur as binding cores for HLA-DRB1*01:01. In this case, NetMHCII predicts that both FQRELNNPL (from tubulin polyglutamylase) and ELTKNAQRA (from C2orf48) form cores within 15-mers that bind with IC_{50} <1000 nmol/L, as shown in **Figure 3.4**B. Based on these predictions, the combination of missense mutation R593C and allele HLA-DRB1*01:01, which at the end of step two (and in earlier computational work [Shepherd et al., 2015]) was predicted to confer a risk of inhibitor development, is ultimately predicted to confers "no, or negligible, risk of inhibitor formation owing to fortuitous cross-matches to peptides in the human proteome" (Hart et al., 2019).

3.3.2 Overview of predicted FVIII inhibitor risk

The calculated inhibitor risk of each distinct combination of HLA allele/missense mutation combination was plotted as a single square on a heatmap (**Figure 3.5** and **Figure 3.6**). The colour of the square represents the highest predicted binding strength of any

risk-associated 15-mers spanning the location of the missense mutation location, with black square indicating that either there is no 15-mer that has that missense mutation within its predicted binding core for that HLA allele, or no such peptides are predicted to form a novel peptide-MHC surface; hence representing low/negligible risk prior to proteome scanning. In **Figure 3.6**, a grey square represents full proteome protection and implies that the predicted status of the relevant FVIII missense mutation/HLA allele combination has changed (owing to the detection of one or more proteome cross-matches) from being associated with the risk of inhibitor formation to being of low or negligible risk. A change to a colour with respect to **Figure 3.5** other than grey implies partial proteome protection – one or more peptides are no longer considered risk associated (owing to the detection of one or more proteome cross-matches), but there remains at least one peptide-MHC surface predicted to be risk associated. Full heatmaps covering all FVIII missense mutations with and without proteome scanning are available in **Appendix 1** (Figures **A2** and **A1** respectively).

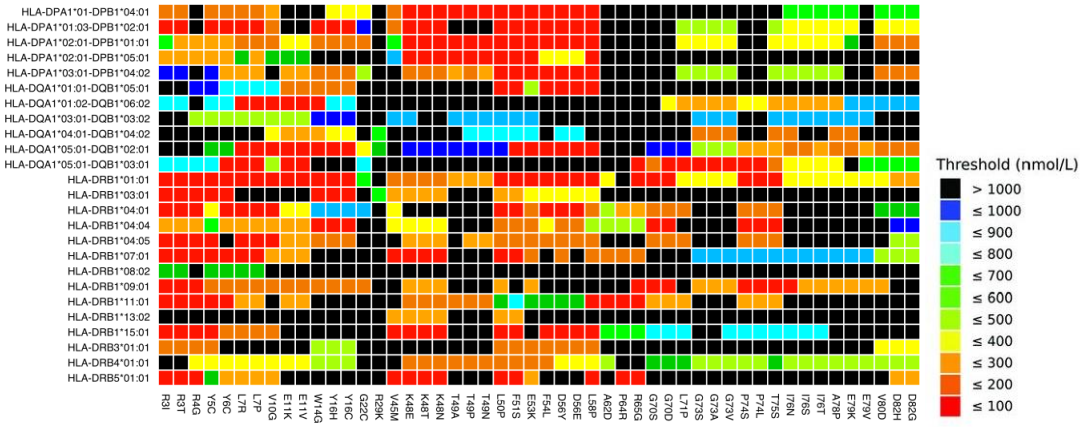


Figure 3.5 Heatmap showing inhibitor risk for a set of missense mutation/HLA allele combinations without proteome scanning (figure reused from [Hart et al., 2019] with permission to reuse the material granted by Ferrata Storti Foundation). The full heatmap for all available missense mutations in the Factor VIII Gene (F8) Variant Database without proteome scanning is available in Appendix 1, Figure A1.

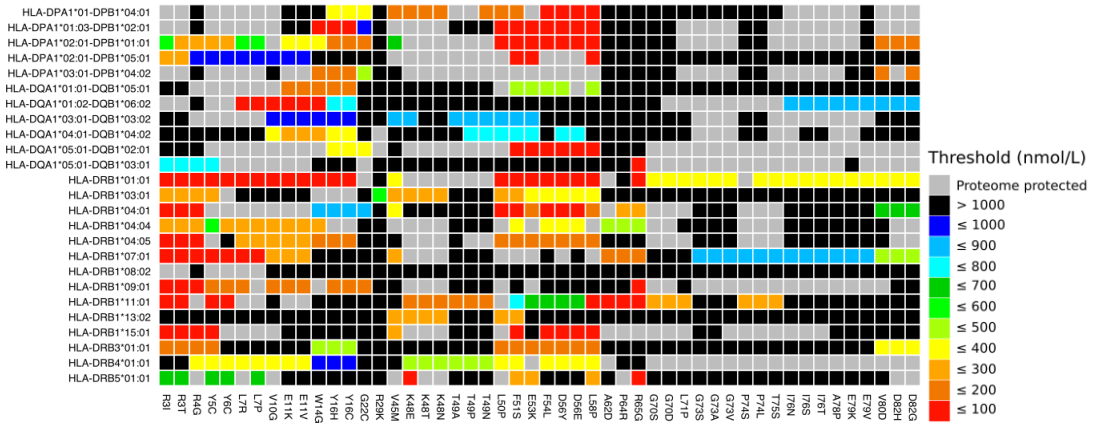


Figure 3.6 Heatmap showing inhibitor risk for a set of missense mutation/HLA allele combinations with proteome scanning (figure reused from [Hart et al., 2019] with permission to reuse the material granted by Ferrata Storti Foundation). Differences between this heatmap and that in Figure 3.5 are attributable to the impact of proteome scanning. The full heatmap for all available missense mutations in the Factor VIII Gene (F8) Variant Database with proteome scanning is available in Appendix 1, Figure A2.

Our overall assessment of the significance of proteome scanning on predicted inhibitor risk was as follows. Taking account of all HLA allele/FVIII missense mutation combinations in the set (25 alleles x 956 mutations), it was predicted that, with a conservative threshold of IC_{50} < 1000 nmol/L, the percentage of inhibitor risk-associated squares falls from 49% to 31% when proteome cross-matches are taken into account. The rate falls from 37% to 21% with a threshold of IC_{50} < 500 nmol/L and from 29% to 15% with a threshold of IC_{50} < 300 nmol/L. Note that, whereas different binding thresholds may be important in the context of inhibitor risk (given that stronger-binding peptides are more likely to induce an immune response), the standard IC_{50} < 1000 nmol/L threshold was consistently retained for calculating the likelihood of self-tolerance with respect to peptides in the human proteome.

The predicted inhibitor risk associated with individual HLA alleles was also calculated and is shown in **Table 3.1**. Clearly the predicted risk for each HLA allele differs significantly, indicating the importance of the HLA type in prediction of inhibitor risk.

Table 3.1 Breakdown of predicted inhibitor risk with respect to different HLA alleles at different thresholds before and after proteome scanning (adapted from [Hart et al., 2019] with permission to reuse the material granted by Ferrata Storti Foundation)

HLA allele	Risk (%) with 1000 nmol/L threshold		Risk (%) with 500 nmol/L threshold		Risk (%) with 300 nmol/L threshold	
	before	after	before	after	before	after
DRB1*01:01	86	48	78	41	71	33
DRB1*03:01	34	25	24	16	19	12
DRB1*04:01	62	38	47	25	38	22
DRB1*04:04	66	38	53	26	44	19
DRB1*04:05	60	35	49	28	39	21
DRB1*07:01	70	46	60	39	47	30
DRB1*08:02	36	23	16	10	8	4
DRB1*09:01	66	41	50	25	40	8
DRB1*11:01	58	37	43	24	38	22
DRB1*13:02	34	25	26	16	19	12
DRB1*15:01	64	36	49	25	37	8
DRB3*01:01	35	25	24	17	18	11
DRB4*01:01	60	33	39	16	30	12
DRB5*01:01	58	40	48	31	40	23
DPA1*01-DPB1*04:01	41	27	33	23	28	18
DPA1*01:03-DPB1*02:01	46	29	36	21	30	8
DPA1*02:01-DPB1*01:01	59	34	47	25	37	20
DPA1*02:01-DPB1*05:01	30	18	17	8	12	6
DPA1*03:01-DPB1*04:02	52	33	41	22	29	12
DQA1*01:01-DQB1*05:01	27	21	19	15	15	10
DQA1*01:02-DQB1*06:02	54	34	36	20	26	16
DQA1*03:01-DQB1*03:02	21	15	10	6	5	3

HLA allele	Risk (%) with 1000 nmol/L threshold		Risk (%) with 500 nmol/L threshold		Risk (%) with 300 nmol/L threshold	
	before	after	before	after	before	after
DQA1*04:01-DQB1*04:02	21	13	11	6	7	4
DQA1*05:01-DQB1*02:01	36	18	24	12	6	7
DQA1*05:01-DQB1*03:01	58	32	40	20	30	13

Additionally, missense mutations associated with "low/negligible risk" independent from the HLA allele were identified; for the IC_{50} < 1000 nmol/L threshold, these correspond to all-black (**Figure 3.5**) or grey-black (**Figure 3.6**) columns in the heatmaps. At this threshold, 25 mutations out of 956 were predicted to be associated with "low/negligible risk" without proteome cross-matches, compared with 40 when proteome cross-matches are taken into account. This number increases with less conservative thresholds, as shown in **Table 3.2**. Nevertheless, it is clear that inhibitor risk is largely HLA dependent.

Table 3.2 Number of missense mutations (from a total of 956) associated with "low/negligible" inhibitor risk

Threshold (nmol/L)	Low risk mutations before proteome scanning	Low risk mutations after proteome scanning
1000	25	40
500	40	80
300	55	154
200	79	209
100	152	353
50	277	492

3.3.3 Analysis of proteome cross-matches

The cross-matches to the human proteome were analysed for all the missense mutation/HLA allele combinations and the set of proteins affording the most proteome protection were identified (see **Table 3.3**). The protein affording the most protection is coagulation factor V, which is known to have a high sequence similarity to FVIII

(Davidson et al., 2003). It is also worth noting that the top four protective proteins – coagulation factor V, hephaestin-like protein 1, ceruoplasmin and hephaestin – all have copper-binding sites, a property they share with FVIII itself.

Table 3.3 List of 15 human proteins affording the highest number of proteome cross-matches (adapted from [Hart et al., 2019] with permission to reuse the material granted by Ferrata Storti Foundation)

Ensembl ID	UniProt ID	Protein Name	Protected peptide count
ENSP00000356771	P12259	Coagulation factor V	640
ENSP00000313699	Q6MZM0	Hephaestin-like protein 1	457
ENSP00000264613	P00450	Ceruloplasmin (ferroxidase)	437
ENSP00000430620	Q9BQS7	Hephaestin	389
ENSP00000353393	P00451	Coagulation Factor VIII (match to different location within the protein)	251
ENSP00000355910	O75445	Usherin	150
ENSP00000431216	Q14585	Zinc finger protein 345	142
ENSP00000444412	Q14587	Zinc finger protein 268	134
ENSP00000254579	Q96M86	Dynein heavy chain domain 1	83
ENSP00000367086	O00154	Acyl-CoA thioesterase 7	76
ENSP00000326563	Q7LBC6	Lysine (K)-specific demethylase 3B	75
ENSP00000358407	Q9UKF2	Disintegrin and metalloproteinase domain-containing protein 30	75
ENSP00000444747	Q9Y2P0	Zinc finger protein 835	74
ENSP00000224600	P10745	Retinol binding protein 3	73
ENSP00000439288	Q5T5N4	Chromosome 6 open reading frame 118	67

The "protected peptide count" for a given human protein aggregates all the protective cross-matches for all combinations of missense mutations and HLA alleles. A single risk-associated FVIII peptide may increment the counts for multiple proteins and increment the count for a single protein by more than one (if the same matching core occurs at multiple locations within that protein). Note that a cross-match for a given peptide is only deemed protective if that peptide is predicted to confer an inhibitor risk before proteome scanning is undertaken.

3.3.4 Evaluation of risk prediction accuracy

The accuracy of the predictions made by the methodology described in preceding sections was evaluated by comparing them to the real patient data from FVIII Gene (F8)

Variant Database. Table 3.4 reports the statistical calculations for missense mutation HA patient data.

Given that there are multiple reasons why an individual that is genuinely at risk of developing inhibitors may not yet have developed them (e.g. if they have had insufficient exposure to tFVIII, or they have developed anti-tFVIII antibodies that are non-neutralising), the most appropriate numerical indicator of the accuracy of our method is the number of false negatives (i.e. the number of patients predicted to have "low/negligible risk" of inhibitor formation who actually developed inhibitors). Table 3.4 shows that the number of false negatives is low at conservative cut-offs (column 3). Using the novel proteome scanning approach, the number of false negatives is somewhat higher (the lower half of column 3), but the total number of patients predicted to be at low/negligible risk is considerably higher. Overall, the proteome scanning predictions have considerably higher statistical significance (final column of Table 3.4).

The higher number of false negatives with proteome scanning implies that some of the cross-matches that are predicted to be protective are not actually protective. There are several reasons why a putative protective cross-match may not be protective in reality: the predicted binding of the cross-matching peptide (either its register or its strength) may be inaccurate; assumptions about one or more HLA anchoring positions may be incorrect (such that a residue mismatch at a presumed anchoring position is in fact TCR-facing); the individual's self-peptide may not match that from the canonical proteome (because it contains a non-synonymous SNP); the cross-match may be to a protein that participates ineffectively in the central tolerance process; or the individual may have one or more comparatively uncommon HLA alleles not considered in the evaluation.

Table 3.4 Fisher's Exact Tests evaluating the accuracy of predicted inhibitor (adapted from [Hart et al., 2019] with permission to reuse the material granted by Ferrata Storti Foundation)

IC_{50} binding threshold (nmol/L)	Patients predicted to have low/negligible risk		Patients predicted to have an inhibitor risk		*P*
	No inhibitors	Inhibitors	Inhibitors	No inhibitors	
Without proteome scanning					
1000	28	1	116	1344	0.72
500	49	3	92	985	0.62
300	122	3	84	787	*5.84-04*
200	179	9	76	660	*0.02*
100	362	20	37	338	*0.02*
50	593	36	31	228	*2.01-03*
With proteome scanning					
1000	103	4	80	622	*0.02*
500	157	7	65	339	*4.50e-05*
300	322	14	57	261	*1.14e-08*
200	465	26	53	232	*1.07e-08*
100	777	42	23	133	*6.57e-05*
50	1114	66	22	115	*3.72e-05*

Column 2 (true-negatives): Patients without inhibitors having a missense mutation predicted to have "low/negligible" risk of inhibitor development
Column 3 (false-negatives): Patients with inhibitors having a missense mutation predicted to have "low/negligible" risk of inhibitor development
Column 4 (true-positives): Patients with inhibitors having a missense mutation predicted to have risk of inhibitor development
Column 5 (false-positives): Patients without inhibitors having a missense mutation predicted to have risk of inhibitor development

3.4 Discussion

The results presented in this chapter have potential clinical relevance in that they reveal the extent to which inhibitor risk is HLA-dependent, and hence (as noted in the paper based on this work) provide "compelling evidence of the importance of HLA class II genotyping for analysing the inhibitor risk of patients with missense mutation hemophilia A" (Hart et al., 2019). This research also represents the first application of a new strategy

termed proteome scanning – the identification of cross-matching peptides within the human proteome as sources of "protection" against adverse immune responses (driven by underlying tolerance mechanisms). The inhibitor risk predictions presented in section 3.3.2 in combination with the statistical analysis of section 3.3.4 demonstrate that proteome scanning offers useful insights (it reduces the predicted level of inhibitor risk development from 49% to 31% using a conservative threshold of $IC_{50} < 1000$ nmol/L) and is sufficiently accurate (using the same threshold, only 4 out of 107 patients predicted to have negligible risk of inhibitor development were reported to have developed them). The potential wider applicability of proteome scanning is addressed in the next chapter.

Notwithstanding these highly promising results, there are several ways in which the method presented here has been simplified in ways that are likely to affect its accuracy, some of which may be addressed in future work. Firstly, the relationship between endogenous and therapeutic FVIII has been simplified in ways that potentially underestimates the number of sequence differences that may trigger a T cell response and ultimately lead to inhibitor formation. An increasing number of patients has been treated using B-domain-deleted (BDD) tFVIII products incorporating novel linker sequences. While these novel linker sequences may be associated with increased inhibitor risk (Sauna et al., 2012), one particular study attempted to assess the immunogenicity of the linker peptides of three BDD recombinant FVIII products via *in vitro* and *in silico* methods and concluded that novel linker sequences of the studied products posed low immunogenicity potential (Bartholdy et al., 2018). And for a relatively small subset of patients, there may be one additional (non-disease associated) mismatch between their endogenous FVIII and tFVIII owing to the presence of uncommon FVIII alleles in the wider population (Viel et al., 2009).

Secondly, only a single proteome is used for proteome scanning purposes. In cases where the proteome of an individual patient is available, a personalised approach to proteome scanning would be possible. However, given that individual genomes vary by only around 0.1%, using a personalised proteome approach is unlikely to change peptide cross-matching for most individuals.

Thirdly, there are several ways in which the relationships between peptides, MHC molecules and TCRs have been simplified in this research. Certain MHC class II molecules have pockets at non-canonical positions (i.e., other than positions 1, 4, 6 and

9), although which HLA alleles fit into this category is unclear, as information about them not systematically documented. Predictions for such MHC molecules would clearly be improved by integrating knowledge of their anchor positions into the prediction workflow. The current model also focuses exclusively on 15-mers, whereas peptides of various lengths may be presented by MHC class II molecules and the impact of different lengths of flanking peptide can be modelled to some degree by tools such as NetMHCII. In the absence of tools capable of predicting peptide cleavage within the class II presentation pathway, the only way to address this point would be to routinely make predictions using peptides of different lengths containing the same core. Whether this would be useful and justify the additional compute time is unclear. It is also worth noting that it is known that, in specific cases, differences in peptide-MHC surfaces are attributable to differences in amino acids at anchoring positions in the MHC groove (Kersh et al., 2001) or outside the binding core (Deng et al., 2007). However, given that the prevalence of such effects is poorly understood, it is unclear how they might be incorporated into a revised computational model.

Finally, it is worth noting that the patient data used in the current work was retrieved from the EAHAD FVIII Gene Variant Database in November 2016. Subsequently, additional data has been added to this database, including patient data from the My Life, Our Future initiative (MLOF) that involved the genotyping of 3000 haemophilia patients (Johnsen et al., 2017). Data is also available (though not via the EAHAD database) for 1112 non-severe haemophilia A patients that was generated as part of the INSIGHT study (Eckhardt et al., 2013). This additional data could be used to update the statistical evaluations presented in this thesis, but the lack of information about the HLA types of patients limits its usefulness. Consider two of the key missense mutations identified in the INSIGHT study: Arg2150His (57 patients, 15.8% inhibitor rate) and Trp2229Cys (10 patients, 50% inhibitor rate). From the heatmap with proteome scanning covering 25 common HLA alleles (**Figure A2**), Arg2150His and Trp2229Cys are predicted to have an inhibitor risk with 8 and 5 alleles respectively. Hence, our prediction of inhibitor risk is consistent with the observed occurrence of patient inhibitors. However, given that we don't know how many of the INSIGHT patients have one or more of the HLA alleles we predict to be risk-associated, there is limited scope for us to make more compelling inferences (e.g. about whether our method successfully predicts the degree of inhibitor risk for a particular missense mutation).

4 Predicting the Risk of Transplant Rejection

4.1 Introduction

4.1.1 Proteome scanning and alloimmunity

In chapter 3, a novel proteome scanning method for assessing inhibitor risk formation was introduced in the context of missense mutation haemophilia A, which is an example of *alloimmunity* – an immune response to a non-self-antigen that originates from the same species, known as an *alloantigen*, commonly involving alloantigen recognition, or *allorecognition*, by T cells. Given this broader perspective on proteome scanning, a natural question to ask is: in what other alloimmune contexts might this or a similar approach prove potentially useful?

As an example of alloimmunity, haemophilia A has several characteristics that make it a fertile target for the proteome scanning approach. Firstly, although haemophilia A is a rare disease, it is well studied and there is a good deal of data in the public domain (see http://www.factorviii-db.org). Secondly, the inhibitor rate among individuals with haemophilia A is up to 30% (Lieuw, 2017), and the prediction of inhibitor risk is potentially useful. Thirdly, the number of differences between endogenous and therapeutic FVIII (tFVIII) may, in many cases of missense mutation haemophilia A, be as few as one. By way of contrast, alternative causes of haemophilia A include large inversions or deletions (Gouw et al., 2012), which imply large differences between endogenous and therapeutic FVIII. In such cases, it is highly likely that novel peptide-MHC surfaces will be presented and that individuals with these mutations will, from a T cell perspective, be at risk of developing inhibitors. In such cases, the proteome scanning approach is unlikely to be very informative.

Dozens of other protein therapeutics are known to trigger an adverse immune response among a proportion of their recipients, although rarely as high a proportion as that associated with tFVIII (Baker et al., 2010). What is less clear is whether any of these diseases a) are (at least) sometimes associated with small differences between endogenous and therapeutic and b) have sufficient available data for predictions to be evaluated.

Bearing these points in mind, a different alloimmune challenge was undertaken for this thesis concerning the risk of transplant rejection.

4.1.2 Transplant rejection: overview

Organ or tissue transplantation is the only treatment option left for some patients with severe and well-advanced illnesses. A within-species transplant is known as an *allograft*. The most important difficulty to overcome in the transplantation process is the possibility of allograft rejection by the recipient. Immunosuppressive medication is commonly used to prevent allograft rejection but often has significant side effects (see, for example, [Moini et al., 2015]). Kidneys are the most commonly transplanted organ globally, mostly from deceased donors; a fairly recent review quotes a 5-year allograft survival rate of just over 70% for deceased donor kidney transplants in the US (Wang et al., 2016).

The most important cause of allograft rejection are MHC molecule mismatches between donor and recipient (Wood & Goto, 2012), although rejection is possible even with perfect HLA matching (e.g., because of minor histocompatibility antigen mismatches [Perreault et al., 1990]). Henceforth, this chapter focuses exclusively on MHC molecule mismatches as the cause of allograft rejection.

T cell-mediated alloimmunity can be triggered via two main mechanisms, known as the direct and indirect pathways. The *direct pathway* involves recipient T cells recognising intact donor MHC molecules presented on the surface of donor antigen presenting cells (APCs). The *indirect pathway* involves recipient T cells recognising peptides derived from donor MHC molecules after they have been internalised, processed and presented on the surface of recipient APCs via the antigen presentation pathways described in sections 1.1.1 and 1.1.2. Given that the number of donor dendritic cells is limited and decreases over time, the direct pathway is associated with short-term, acute rejection. The indirect pathway, on the other hand, is associated with long-term, chronic rejection. (A third pathway –the semi-direct pathway – has also been proposed whereby donor MHC molecules are internalised by recipient APCs and are presented intact on the cell surface [Herrera et al., 2004].)

Although both $CD4^+$ and $CD8^+$ T cells are involved in the direct pathway, $CD4^+$ T cells dominate the indirect pathway, and facilitate the development of anti-graft antibodies (alloantibodies). However, there is "limited but intriguing" evidence that $CD8^+$ T cell

allograft rejection can also be associated with the indirect pathway via "cross-priming" (Lin & Gill, 2016), with skin allograft rejection as a notable example (Valujskikh et al., 2002). Both $CD4^+$ and $CD8^+$ T cell responses are considered in this research.

In the context of bone marrow and stem cell transplants, an additional complication may occur whereby T cells within the allograft (marrow or stem cells) attack the host, and in particular the host MHC molecules, causing a condition known as graft versus host disease (GVHD). GVHD occurs in two forms – acute and chronic – that are immunologically distinct and involve different T cell subsets. For a recent and thorough review, see (Hill et al., 2021).

4.1.3 HLA matching strategies

As explained in section 1.1.4, an individual has several different kinds of MHC class I and class II molecules, is likely to exhibit high levels of heterozygosity with respect to these molecules, and the number of other individuals sharing the same MHC molecules is likely to be small because the HLA genes that encode these molecules are highly polymorphic. Consequently, perfect HLA matching between donor and recipient is rarely possible. In practice, the desirability of a close HLA match (with potential benefits in terms of allograft and/or patient survival and reduced immunosuppression) has to be balanced against the impact of longer waiting times (Zachary & Leffell, 2016). The optimal degree of donor/recipient HLA matching is context dependent. For example, exact allelic matching at (at least) four loci (HLA-A, -B, -C and -DRB1) has generally been the preferred option in the context of bone marrow transplantation (Lee et al., 2007); in the latter case, HLA matching is additionally important with respect to GVHD (Loiseau et al., 2007).

But what constitutes an HLA match? Until comparatively recently, HLA typing methods have focused on the most variable region of MHC molecules, the peptide binding groove (Erlich, 2012), thereby ignoring variation in other parts of the molecule. In 2019, an "ultra-high resolution" sequencing method – one that uses Pacific Biosciences Single Molecule Real-Time sequencing (PacBio) as a way of overcoming the ambiguities that arise with standard short-read Next Generation Sequencing – for HLA-A, -B, -C, -DRB1, -DQB1 and DPB1 was published (Mayor et al., 2019), but this has yet to be widely adopted. PacBio sequencing is a third-generation sequencing method offering major

improvements over the common problems and limitations of the second-generation sequencing (SGS) methods. For example, short read lengths require the usage of multiple overlapping sequences that makes the SGS methods vulnerable to incorrect alignments and consequently to HLA allele typing errors. As the HLA gene region is one of the most polymorphic loci in the human genome, incorrect phase resolution of the polymorphism may preclude identifying the correct allele assignment. PacBio enables HLA sequencing to be achieved in a single reaction with longer reads, allowing accurate isoform detection and more reliable allele assignments (Mayor et al., 2015).

Irrespective of the resolution of matches, the simplest strategy for measuring the degree of mismatch is to count the number of mismatching residues between the MHC molecules of the donor and those of the potential recipient. In effect, all mismatches are treated equally. Alternatively, individual mismatches may be considered "permissive" or "nonpermissive" depending on their observed or predicted impact on transplantation outcomes. The most notable example of this approach involves mismatches in the HLA-DPB1 locus; predictive algorithms have been developed that assign mismatches to the permissive or nonpermissive class according to whether the mismatched DPB1 molecules are functionally similar in terms of the T cell epitopes they present (Meurer et al., 2021; Zino et al., 2004).

A completely different strategy for identifying permissive mismatches has been developed that is based on a structural appraisal of the location of MHC residue mismatches: if they are at exposed positions on the intact MHC molecule, they are likely to be detected as non-self by alloantibodies and hence should be considered nonpermissive. The HLAMatchmaker algorithm that implements this approach originally focused on sequence triplets (in effect short, linear B cell epitopes) (Duquesnoy, 2002), but subsequently accommodated potential discontinuous B cell epitopes by defining a 3-3.5 Å radius around a given surface location (Duquesnoy, 2006).

4.1.4 A proteome scanning-based strategy for the detection of permissive mismatches

The method presented here takes a different approach to the identification permissive and nonpermissive mismatches based on the strategy for identifying novel peptide-MHC surfaces described in section 3.2.1 and the proteome scanning strategy described in 3.2.2.

The basic rationale behind this approach is that certain mismatches can be labelled permissive because they are effectively invisible to the immune system because a) they do not occur within the binding core of a T cell epitope, or b) they occur in a binding core, but not at a TCR-facing position, or c) they cross-match to peptides elsewhere in the human proteome.

4.2 Methods

4.2.1 Modifications to the proteome scanning approach

To implement this approach, an important modification to proteome scanning approach was required to ensure that the correct MHC molecules (and only the correct MHC molecules) were incorporated in the reference proteome. To achieve this, all MHC molecules were removed from the Ensembl reference proteome (see section 3.2.2) and the correct sequences added on a case-by-case basis as appropriate – either the transplant recipient's MHC sequences or, in the context of a GVHD evaluation, the donor's MHC sequences.

The most recent versions (in 2019) of the Technical University of Denmark (DTU) tools were used in this research: NetMHCIIpan for MHC class II binding prediction and NetCTLpan for combined MHC class I C-terminal proteasomal cleavage, TAP transport efficient and binding prediction (see section 2.2).

4.2.2 The selection of examples for method valuation

Unfortunately, there is a lack of data in the public domain about rejection rates associated with specific HLA mismatches[2]. For this research, a single published case study became the key focus for analysis: a case of bone marrow allograft rejection where donor and recipient were perfectly HLA matches except for a single amino-acid residue difference in HLA-B*44 – the donor had allele HLA-B*44:03 with a Leucine (L) at position 180 (numbered 156 in the original paper), and the recipient allele HLA-B*44:02 with an

[2] To remedy this problem, a collaboration (covered by a non-disclosure agreement) was initiated that would have facilitated access to a large set of patient data containing information about their HLA allelic mismatches and rejection status. Unfortunately, progress with this collaboration has not, as yet, taken place because of Covid.

Aspartic Acid (D) at the same position. This represents a minimal mismatch using a "default" strategy of residue-mismatch counting. The recipient died 68 days after transplantation, with "allocytotoxic host-derived $CD8^+$ T cells [observed] in the patient's circulation at the time of rejection" (Fleischhauer et al., 1990).

In the paper describing this case study, shared HLA alleles are listed as HLA-A2, -A3, -B7, -B44, -DR2 and -DR7, but this is not in full; most class II alleles are absent, and none of the alleles are specified at full resolution. To assess the impact of this specific HLA-B*44 mismatch on potential novel peptide-MHC surface formation for a) the individual recipient in the case study and b) other potential donors/recipients with the same, singular mismatch, the following analyses were undertaken. For MHC class II, the set of 25 alleles with wide population coverage that was used in the haemophilia A research (see section 3.2.1) were again used here. For MHC class I, two approaches were adopted. Firstly, the most common allelic variants of the listed HLAs – HLA-A*02, HLA-A*03 and HLA-B*07 – were tested (e.g. HLA-A*02:01 and HLA-A*02:02 for HLA-A*02) based on the population frequencies recorded in the Allele Frequency Net Database (see section 2.1). Secondly, a panel of 20 common HLA alleles with wide population coverage were compiled from the Allele Frequency Net Database and used to give a broader insight into the potential risks associated with this HLA-B*44 mismatch.

4.3 Results

*4.3.1 The HLA-B*44:03 (donor) vs. HLA-B*44:02 (recipient) mismatch*

Given that $CD8^+$ T cells were observed in the recipient at the time of allograft rejection (section 4.2.2) and his HLA class I alleles were at least partially specified, this is the natural place to start the analysis. The outcome is shown in **Figure 4.1**. The allograft recipient was known to have the HLA-B*44:02 allele together with alleles from HLA-A*02, HLA-A*03 and HLA-B*07. Common alleles from the latter three were chosen (2 for HLA-A*02, 2 for HLA-A*03 and 1 for HLA-B*07) based on their global population frequencies. The colour of a square represents the highest binding affinity of any 9-mer spanning location 180 within the HLA-B*43:03 sequence that has location 180 at a TCR-facing position. Black implies there is no such binding peptide with a binding affinity ≤ 1,000 nmol/L.

These results suggest that the recipient was likely to have had an HLA-A*02 allele capable of initiating a $CD8^+$ T cell response to the allograft. In this instance, proteome scanning (**Figure 4.1**B) made no difference. It is worth noting that, MHC class I-bound peptides are less likely to cross-match to the human proteome than MHC class II-bound peptides because the former typically have 7 TCR-facing residues (given canonical class I anchoring positions 2 and 9) whereas the latter typically have only 5 TCR-facing residues (given canonical class II anchoring positions 1,4, 6 and 9).

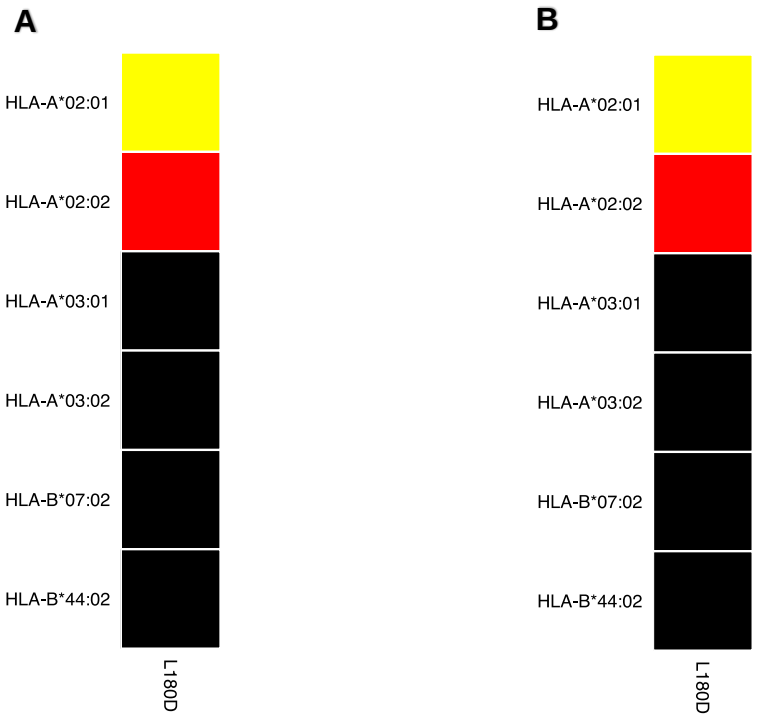


Figure 4.1 Heatmap showing the immunogenicity of the HLA-B*44:03 (donor) vs. HLA-B*44:02 (recipient) mismatches for recipient HLA class I alleles. (For a temperature key, see Figure 3.5.) A) Without proteome scanning. B) With proteome scanning.

Given that the allograft recipient's HLA class II alleles were unspecified, an equivalent evaluation was carried out using the set of 25 common HLA class II alleles previously used in Chapter 3. The outcome is shown in **Figure 4.2**. The colour of a square represents the highest binding affinity of any 15-mer spanning location 180 within the HLA-B*43:03 sequence that has location 180 at a TCR-facing position. Black implies there is no such binding peptide with a binding affinity ≤ 1,000 nmol/L. A grey square implies that all relevant peptides (i.e., those that bind to the relevant MHC molecule with an affinity ≤ 1,000 nmol/L and form a novel peptide-MHC surface) cross-match to

binding peptides in the proteome. A change to a colour other than grey implies the peptide(s) with the strongest binding affinity and forming a novel peptide-MHC surface cross-match to binding peptides in the proteome, but there remains at least one such peptide with a binding affinity ≤ 1,000 nmol/L. These results suggest that the recipient was rather unlikely to have had an HLA class II allele capable of initiating a $CD4^+$ T cell response to the allograft – but only when cross-matches to the human proteome (using the novel proteome scanning method) are taken into account, as shown in **Figure 4.2**B.

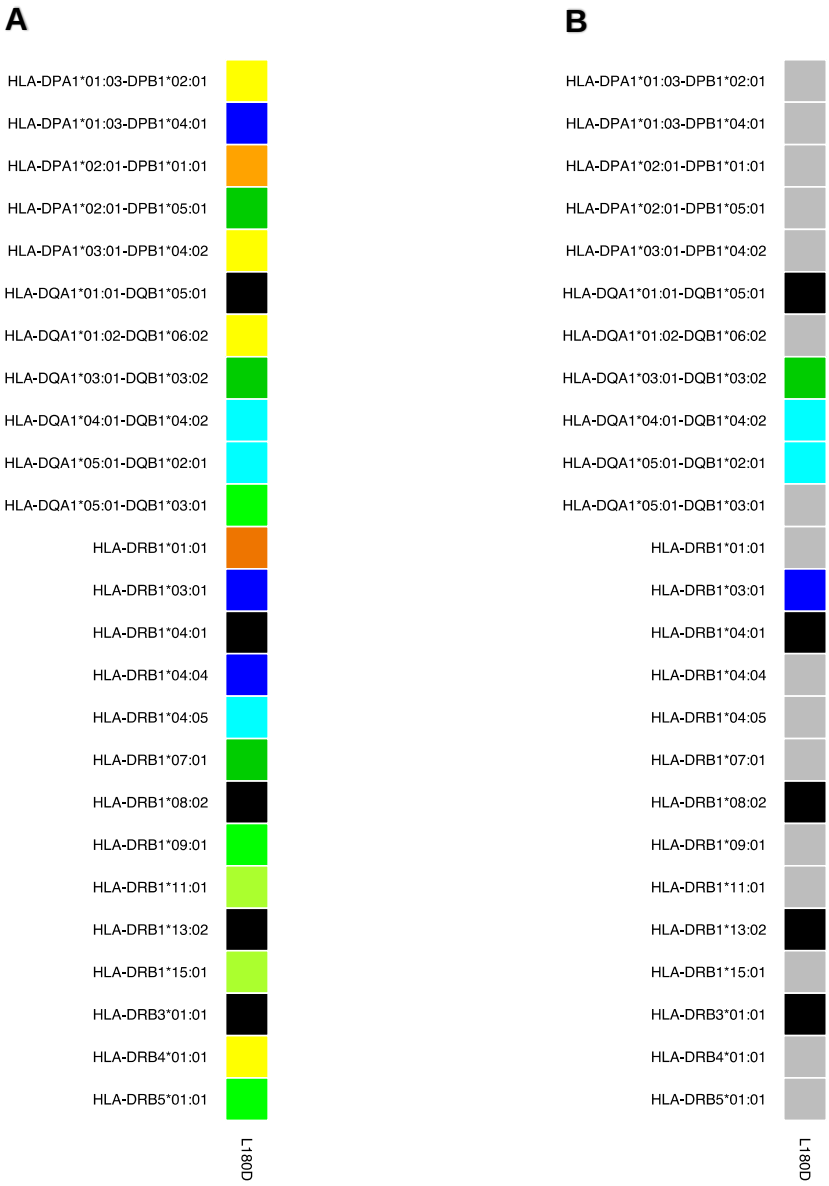


Figure 4.2 Heatmap showing the immunogenicity of the HLA-B*44:03 (donor) vs. HLA-B*44:02 (recipient) mismatches for a set of 25 common HLA class II alleles. The colour of a square represents the highest binding affinity of any 15-mer spanning location 180 within the HLA-B*43:03 sequence that has

location 180 at a TCR-facing position. Black implies there is no such binding peptide with a binding affinity ≤ 1,000 nmol/L. (For a temperature key, see Figure 3.6.) A) Without proteome scanning. B) With proteome scanning.

*4.3.2 Mismatch counts for different HLA-B*44 alleles*

To put the HLA-B*44:03 vs. HLA-B*44:02 mismatch in wider context, the number of pairwise mismatches between four HLA-B*44 alleles was calculated together with the corresponding assessment of mismatch "permissiveness" using the proteome scanning approach. The number of pairwise mismatches is shown in **Table 4.1**; each number in a cell represents the count of residue differences between the corresponding HLA sequences in the row and the column. The number of HLA class I alleles (from a set of 20 common alleles) associated with the risk of transplant rejection are shown in **Table 4.2**. and the number of HLA class II alleles (from a set of 25 common alleles) associated with the risk of transplant rejection are shown in **Table 4.3**. Note that the additional proteome scanning step made no difference with respect to the class I predictions (**Table 4.2**), for the reasons given in section 4.3.1.

Table 4.1 Pairwise mismatch counts for HLA-B*44 alleles

Donor (down)	HLA-B*44:02	HLA-B*44:03	HLA-B*44:05	HLA-B*44:04
HLA-B*44:02	-	1	1	2
HLA-B*44:03	1	-	2	2
HLA-B*44:05	1	2	-	3
HLA-B*44:04	2	2	3	-

Table 4.2 Number of HLA class I alleles (from a list of 20 common alleles) associated with a predicted rejection risk given mismatching HLA-B*44 alleles between donor and recipient

Donor (down)	HLA-B*44:02	HLA-B*44:03	HLA-B*44:05	HLA-B*44:04
HLA-B*44:02	-	3	0	5
HLA-B*44:03	7	-	7	9
HLA-B*44:05	2	4	-	6
HLA-B*44:04	5	5	5	-

Table 4.3 Number of HLA class II alleles (from a list of 25 common alleles) associated with a predicted rejection risk given mismatching HLA-B*44 alleles between donor and recipient

Donor (down)	HLA-B*44:02	HLA-B*44:03	HLA-B*44:05	HLA-B*44:04
HLA-B*44:02	-	3 (1)	0	21 (16)
HLA-B*44:03	20 (4)	-	20 (4)	22 (20)
HLA-B*44:05	3 (2)	6 (3)	-	22 (18)
HLA-B*44:04	10 (3)	10 (3)	10 (3)	-

The number of risk-associated alleles after proteome scanning is given in brackets.

The most important feature of this analysis is that the number of mismatches (**Table 4.1**) is poorly correlated with the predicted rejection risk aggregated across different HLA alleles (**Table 4.2** and **Table 4.3**). For example, the HLA-B*44:03 (donor) vs. HLA-B*44:02 (recipient) mismatch that has been the main focus of this chapter involves only a single sequence mismatch, but this is associated with a predicted rejection risk with 7 out of 20 common HLA-I alleles (**Table 4.2**), whereas the HLA-B*44:04 (donor) vs. HLA-B*44:05 (recipient) involves three sequence mismatches (i.e. three times as many), yet the predicted rejection risk is lower (only 5 out of 20 common HLA-I alleles). In the case of HLA-II alleles (**Table 4.3**), there is an even more striking example: the HLA-B*44:03 (donor) vs. HLA-B*44:04 (recipient) mismatch involving two sequence mismatches has a

high predicted rejection risk (20 out of 25 common HLA-II alleles) whereas the HLA-B*44:04 (donor) vs. HLA-B*44:05 (recipient) mismatch involves an additional sequence mismatch (i.e. three residues) but a much lower predicted rejection risk (only 3 out of 25 common HLA-II alleles).

4.4 Discussion

In this chapter, a single HLA allele mismatch known to be associated with bone marrow allograft rejection has been analysed using a modified version of the proteome scanning approach presented in Chapter 3. This narrow focus, which reflects the lack of publicly available data, is an obvious limitation, and one that (through appropriate collaboration) there is scope to remedy in the near future.

Nevertheless, the results presented here suggest there is considerable scope for the computational prediction of transplant rejection risk using an assessment of whether a given mismatch will be visible to the recipient's MHC molecules – or, in the case of GVHD, the donor's MHC molecules. Such an approach would be particularly timely given that the HLA typing of donors and recipients is routinely undertaken and increasingly with high resolution (in clear contrast with haemophilia A patients, who are rarely HLA typed). In principle, such an approach could make a two-fold contribution: by identifying high-risk mismatches, transplants that are likely to result in early rejection may be prevented; and by identifying low-risk mismatches, the pool of potential donors for a given recipient could be expanded.

Ultimately, as in the haemophilia A case, the accuracy of this new computational method is a crucial issue, and one we will return to in the Conclusion of this thesis. The main obstacle in the context of our work on transplant rejection prediction is the lack of available data. The priority, therefore, to facilitate further progress with this research is to gain access to a dataset containing detailed information about many (preferably hundreds of) transplant patients. Such data should contain a high-resolution specification of donor and recipient HLA types, both class I and class II, together with an assessment of each donor's post-transplantation outcomes. Such data exists, but is not currently available in the public domain.

5 T cell epitopes and the detection of anti-tumour immunity in HCC

5.1 Introduction

In the research presented within this chapter, the methods previously applied to identify the peptides associated with the risk of Factor VIII inhibitor development (Chapter 3) and transplant rejection (Chapter 4) are used in the detection of anti-cancer immunity in the specific context of hepatocellular carcinoma (HCC) as part of a collaboration with researchers at the Institute of Hepatology. Before discussing HCC, it is worth introducing the topic of cancer immunotherapy more broadly.

5.1.1 Cancer immunotherapy: an overview

Cancer immunotherapy utilises a patient's own immune system by stimulating their immune cells or inhibiting certain suppressive pathways to control and eliminate tumours. In 1891, orthopaedic surgeon Dr. William Coley attempted to treat tumours by injecting live and inactivated bacteria into them based on his observation that patients who developed serious wound infections after bone cancer surgery showed regression in the remaining tumour mass. This is considered the first attempt at immunotherapy against cancer (Esfahani et al., 2020). Since that time, researchers have encountered many obstacles in the design and application of effective cancer immunotherapy protocols, and significant progress has arguably only been made in the 21st century, and most particularly in the past decade.

The steps required to establish an effective and long-lasting anti-cancer immunity in a patient is broadly similar to the immune response against pathogens. The process typically begins with dendritic cells capturing and processing antigens that originate from the tumour. Dendritic cells can present tumour antigens via both the class I and class II antigen presentation pathways (sections 1.1.1 and 1.1.2) and activate both $CD4^+$ and $CD8^+$ T cells that are tumour-antigen specific (Gardner & Ruffell, 2016). To be effective, activated tumour-specific T cells commonly need to overcome immunosuppressive mechanisms in the tumour microenvironment, with implications for anti-cancer immunotherapies (for a recent and detailed review, see [Labani-Motlagh et al., 2020]).

In spite of the complexities and challenges, immunotherapy has become a viable treatment option for cancer alongside classical modalities such as surgery, radiotherapy and chemotherapy. By 2019, the U.S. Food and Drug Administration (FDA) approved immunotherapies for patients for over 20 cancer types (www.cancerresearch.org). This includes cancer vaccines for the treatment of early-stage bladder cancer (TheraCys and TICE), metastatic castration-resistant prostate cancer (PROVENGE) and metastatic melanoma (IMLYGIC) (DeMaria & Bilusic, 2019).

Cancer immunotherapy methods can be classified into two groups – active or passive – based on their status as stimulators of the host's own immune cells. Active therapies attempt to stimulate the host immune system and include cancer peptide vaccines that contain fragments of tumour-associated antigens, dendritic cell-based therapies, immunostimulatory cytokines and checkpoint inhibitors. Passive therapies, on the other hand, supply the relevant immune system components (i.e., without relying on the host immune system) and include tumour-targeting monoclonal antibodies (mAbs) and the adoptive transfer of T cells (ACT). The latter includes: tumour-infiltrated lymphocytes (TIL) therapy, whereby unmodified tumour-infiltrating autologous T cells are isolated, expanded *ex vivo*, then infused back into the patient; TCR gene therapy, whereby autologous T cells are modified via the introduction of tumour-targeting TCRs; and chimeric antigen receptor (CAR) T cell therapy, whereby autologous T cells are modified such that their TCRs incorporate a single-chain variable fragment (scFv) from a tumour-targeting mAb (Rohaan et al., 2019).

To develop effective cancer immunotherapies, the ability to detect and characterise anti-tumour immune responses is crucial, but far from straightforward. Anti-tumour T cells commonly form against two types of tumour antigen: tumour-specific antigens (TSAs) that are expressed exclusively within tumour cells, but which (owing to a property known as tumour heterogeneity [Reardon & Wen, 2015]) may vary within a single tumour, between different tumours in the same individual, and between individuals; and tumour-associated antigens (TAAs) that are expressed (often overexpressed [Bright et al., 2014]) by certain tumour cells, but also by certain non-tumour cells. The mutations associated with TSAs may lead to the presentation of peptides that are novel – known as neoepitopes – capable of inducing a T cell response. TSAs, however, may be susceptible to self-tolerance, although this may be incomplete (see, for example, [Cloosen et al., 2007]).

In many respects, antigenic peptides have a clear therapeutic advantage over whole antigenic proteins: peptides from multiple antigens can be synthesised much more cheaply and used in combination. Antigenic peptides can also be used in therapeutic cancer vaccines, and in this context have the added advantage that they may induce immune responses that are more focused on key (neo)epitope targets. However, a peptide-based approach raises additional challenges: appropriate sets of HLA alleles need to be identified that will give good population coverage and a corresponding set of T cell epitopes capable of binding to these HLA alleles needs to be identified, or predicted (e.g., using the tools described in section 2.2) (Kumai et al., 2017).

The optimal selection of antigenic peptides, whether for identifying anti-tumour immune activity or for vaccination, depends on the type of cancer and the therapeutic context. A strategy for addressing this challenge in the context of hepatocellular carcinoma (HCC) at the Institute of Hepatology will be described in due course.

5.1.2 Hepatocellular carcinoma (HCC)

Hepatocellular carcinoma (HCC) is the most common type of primary liver cancer accounting for 80-90% of cases. According to Cancer Research UK, around 6,100 people are diagnosed with liver cancer every year in the UK. Although not one of the most common types of cancer, accounting for around 2% of all new cases, it is the 8th most common cause of cancer death in the UK; only around 13% of people diagnosed with the disease are expected to survive for 5 years or more.

The liver environment is intrinsically immunosuppressive and tolerogenic as it is exposed to a large and varied range of antigenic products, associated with food and pathogens, transported from the gut via the portal vein. Liver sinusoidal endothelial cells (LSECs) have APC functions, and express both MHC class I and class II molecules. However, LSECs also express high levels of inhibitory molecule PD-L1 and low levels of co-stimulatory molecule CD80 and CD86, and thus create an unsuitable environment for the activation of T cells. Owing to secretion of cytokines such as IL-10 and TGF-β1, MHC expression is also downregulated in these cells. Liver-specific macrophages, known as Kupffer cells, play a similar role in inducing tolerance in the liver microenvironment by secreting immunosuppressive cytokines and facilitating Treg proliferation. Overall, the liver has a large population of $CD8^+$ T cells but relatively smaller population of $CD4^+$ T

cells, natural killer (NK) cells and natural killer T (NKT) cells. Functionally, these immunosuppressive and tolerogenic characteristics are important, as they help prevent liver damage in the context of constant antigen exposure. However, these same characteristics may delay effective anti-tumour immune responses (Crispe, 2011; Ringelhan et al., 2018).

Hepatocellular carcinoma (HCC) is the commonest form of liver cancer and is associated with hepatocytes, the liver's main functional cell type. The development of HCC is closely linked to chronic liver inflammation, fibrosis formation and cirrhosis. Although alcohol related liver disease, obesity and diabetes are HCC-associated risk factors, chronic hepatitis B virus (HBV) infection is the leading cause of HCC worldwide and hepatitis C virus (HCV) infection the leading cause in Western countries (Ringelhan et al., 2018).

These characteristics of HCC may further reduce the effectiveness of anti-tumour immune responses. Chronic liver inflammation promotes T cell exhaustion. $CD8^+$ T cells derived from patients with chronic HBV and HCV infections show high expression levels of CTLA-4 (CD152) and PD-1 (CD279), both of which downregulate immune responses. And $CD4^+$ and $CD8^+$ T cells derived from HCC tumours commonly have diminished functionality (Harding et al., 2016).

Surgical resection, ablation, transplantation and transarterial chemoembolisation are proven treatment options for early to intermediate stage HCC. Tyrosine kinase inhibitors such as serofenib are approved for treating patients with advanced-stage disease, but the survival benefits are comparatively modest (i.e., weeks or months) (Forner et al., 2018).

5.1.3 HCC Immunotherapy

Immunotherapy currently represents the most promising strategy for treating HCC. Given recent advances in checkpoint blockade therapy for the treatment of various cancer types, many HCC patients in the advanced stages of the disease are now being offered checkpoint inhibitor treatment as a first or second line of treatment following the 2017 FDA approval of nivolumab (anti-PD-1).

The most extensively studied immune checkpoint molecules are PD-1 and CTLA-4, both of which have important roles in constraining the T cell response and preventing autoimmunity under normal physiological conditions. CTLA-4 expression is upregulated

in T cells upon TCR binding to peptide-MHC, and CTLA-4 competes with the co-stimulatory molecule CD28 to bind CD80 and CD86 ligands on the APC surface (Wei et al., 2018). PD-1, which is expressed by activated T cells, inhibits T cell activation via binding to its ligands, PD-L1 and PD-L2. Recent research has also shown that PD-1, like CDLA-4, interferes with the CD28 co-stimulatory-signalling pathway (via the recruitment of a phosphatase signalling molecule upon binding) (Hui et al., 2017).

CTLA-4 inhibiting antibody therapy prevents CTLA-4 competing with CD28 for binding to CD80 and CD86 on APCs, thereby promoting the activation of $CD8^+$ T cell and specific $CD4^+$ T cell subsets within lymph nodes. PD-1 blockade therapy mainly acts to revive the effector function of exhausted tumour-specific $CD8^+$ T cells (the impact on the $CD4^+$ T cell response remains unclear). PD-L1 blocking therapy is broadly similar, but may have an additional mode of action involving antibody-dependent cellular toxicity (ADCC) (Wei et al., 2018).

There are four FDA-approved checkpoint inhibitor therapies for HCC. Nivolumab, an anti-PD-1 antibody, was approved in 2017 as a second-line therapy for patients treated with sorafenib with unresectable HCC (El-Khoueiry et al., 2017). A recent clinical phase III study (CheckMate 459) of nivolumab versus sorafenib as first line treatment in patients with advanced HCC failed to achieve predefined statistical significance for overall survival (OS) but nevertheless demonstrated some clinically meaningful improvements in OS (Yau et al., 2019). Subsequent approval has been granted for pembrolizumab (an anti-PD1 antibody), ipilimumab (an anti-CTLA-4 inhibitor) and atezolizumab (an anti-PD-L1 antibody). In addition to the approved drugs, durvalumab (anti-PD-L1) and tremelimumab (anti-CTLA-4) have been granted orphan drug designation and a phase III clinical trial evaluating their combination for first-line treatment is still active. Many other clinical trials of checkpoint inhibitors are underway as mono or combination therapy for HCC – see (Nakano et al., 2020).

A different immunotherapeutic approach for HCC is cancer vaccination. The dominant strategy here is to utilize HCC TAAs with the aim of increasing tumour-specific T cell responses. There are several recognized HCC-associated antigens. α-fetoprotein (AFP) is expressed during foetal development, but its expression falls to very low levels shortly after birth. AFP is commonly reactivated in HCC patients. Many AFP-specific class I epitopes have been identified (Breous & Thimme, 2011). A phase I clinical trial of AFP-

derived vaccine demonstrated a complete response in one patient and stable disease in eight out of 15 patients (Nakagawa et al., 2017).

Glypican-3 (GPC3) is a foetal oncoprotein that is overexpressed in HCC on the surface of HCC cells. GPC3 derived peptides have been evaluated in multiple phase I and phase II clinical trials involving either vaccines or CAR-T cell therapy. In a phase I trial with 33 HCC patients, GPC3 peptide vaccine induced GPC3-specific CTL responses in 30 patients, where CTL frequency correlated with overall survival (Sawada et al., 2012).

Melanoma-associated antigen gene A (MAGE-A) is a family of cancer-testis (CT) antigens expressed in germ cells and various cancers, including HCC. In one study, MAGE-A1 and MAGE-A3 specific tumour $CD8^+$ T cells were shown to exist in HCC patients and these were successfully induced by MAGE-derived epitopes *in vitro* (Zerbini et al., 2004). In another study, researchers evaluated the expression of several TAAs in the HCC tissues from 142 patients and showed that MAGE-A3 and MAGE-A4 expression was correlated with serum AFP, one of the most widely-used early diagnosis and monitoring marker for HCC (M. Wang et al., 2015).

Cancer/testis antigen 1B (NYESO-1) is another CT antigens. Flecken et al. (2014) studied the $CD8^+$ T cell responses to specific peptides from various TAAs, including NYESO-1, in 96 HCC patients, and detected IFN-γ producing $CD8^+$ T cells for all TAAs after *in vitro* expansion and antigen-specific stimulation of T cells (Flecken et al., 2014).

Baculoviral IAP repeat-containing protein 5 (survivin) is an apoptosis inhibitor. It is expressed during foetal development, but its expression is largely repressed in cells during normal development and termination. Survivin expression is upregulated in various tumours including HCC, with expression level correlated with tumour cell proliferation, to unsatisfactory responses to chemotherapy and radiotherapy, and ultimately to poor prognosis (Su, 2016).

Cellular tumour antigen p53 is a transcription factor that acts as a tumour suppressor, notably via its role in regulating the cell cycle and activating DNA repair. Around half of cancers, including HCC, involve mutations in p53. Although wild type p53 is still expressed in tumours, its function may be inhibited by the mutant p53 (Vousden & Lane, 2007).

5.1.4 Context: research at the Institute of Hepatology

The research presented in this chapter is part of a collaboration with wet-lab scientists at the Institute of Hepatology led by Dr. Shilpa Chokshi. The aim of the research was to utilize peptides, predominantly from HCC tumour-associated antigens, to stimulate T cells and NK cells, to monitor immune responses, and to determine whether blockading the PD-1/PD-L1 pathway can stimulate and restore dysfunctional T cells in the presence of selected antigenic peptides. The set of 8 HCC tumour-associated antigens chosen by Dr. Chokshi were as follows: Alpha-fetoprotein (AFP), Glypican 3 (jglyp-3), Melanoma-associated antigen 1 (MAGE1), Melanoma-associated antigen 3 (MAGE3), Melanoma-associated antigen 4 (MAGE4), Cancer/testis antigen 1B (NY-ESO1), Cellular tumour antigen p53 and Baculoviral IAP repeat-containing protein 5 (survivin).

My own contribution was to select an appropriate set of antigenic peptides, as described in the Methods (section 5.2).

5.2 Methods

5.2.1 HLA class I allele selection

The selection of HLA class I alleles was confined to HLA-A and HLA-B, as these were considered of prime importance in the immune response to HCC (Dr. Shilpa Chokshi, personal communication). The objective was to select HLA alleles that give high levels of population coverage at both loci, but proceeded in two stages owing to a change in the target patient cohort – initially Bulgarian nationals (from whom the original HCC tumour samples were acquired), but this was subsequently extended to allow for samples from any major regional population.

As the single Bulgarian sample available in the Allele Frequencies Net Database (accessed in February 2015) contained only 55 individuals, the decision was taken to base the initial selection of alleles on a Romanian sample with HLA-A and HLA-B data for around 6,000 individuals (http://www.allelefrequencies.net/pop6001c.asp?pop_id=2259), Romania being in close geographical proximity and sharing ethnic similarities with Bulgaria. Population coverages of 98% and 96% were achieved for HLA-A and HLA-B alleles respectively.

Regional allele frequencies were taken from the NCBI's dbMHC database for Europe, South-East Asia, South-West Asia and Sub-Saharan Africa. dbMHC is no longer maintained (an archive of dbMHC data is available here: https://ftp.ncbi.nlm.nih.gov/pub/mhc/mhc/Final%20Archive/). A population coverage of greater than 90% was achieved for both HLA-A and HLA-B for all regions.

The final selection of HLA alleles was as follows (alleles are listed in the order in which they were added):

HLA-A: 02:01, 01:01. 03:01, 24:02, 11:01. 29:02. 32:01, 68:01, 31:01, 26:01. 25:01, 23:01, 68:02, 30:02, 30:01, 02:02, 74:01, 36:01, 33:03 (total 19).

HLA-B: 07:02, 08:01, 44:02, 35:01, 51:01. 40:01, 44:03, 15:01, 18:01, 57:01, 14:02, 27:05, 13:02, 38:01, 55:01, 37:01, 35:03, 14:01, 49:01, 50:01, 39:01, 40:02, 53:01, 15:03, 42:01, 58:02, 58:01, 52:01, 78:01, 41:01, 56:01 (total 31).

5.2.2 Selection of tumour-specific epitopes

Although HCC tumours are known to be heterogeneous in many patients (Craig et al., 2020), the COSMIC database (https://cancer.sanger.ac.uk/cosmic) (Forbes et al., 2017) was consulted (February 2015) to identify any mutations that have been observed to occur with relatively high frequency in the tumours of HCC patients.

A single substitution, Arg249Ser, in tumour antigen p53 was observed in 231 unique HCC samples representing a frequency of 0.25 (COSMIC Genomic Mutation ID COSV52661594, COMSIC Legacy Identifier COSM10817). Interestingly, the current (12th May 2022) number of unique samples containing this mutation is 374, with a frequency of 0.60. This mutation occurred rarely as a non-HCC natural variant (frequency < 0.03) within the 1000 Genomes Project catalogue (Via et al., 2010).

Predictions were made for 9-mers spanning the p53 Arg249Ser substitution (based on UniProt sequence P04637) using The MHC class I binding prediction tool NetMHCpan 2.4 (Nielsen et al., 2007; Hoof et al., 2009) together with the tool NetCTLpan 1.1 (Stranzl et al., 2010), which makes combined predictions for proteasomal cleavage, TAP transport and MHC class I binding). In both cases, a binding threshold of $IC_{50} \leq 500$ nmol/L was chosen. (Both tools are discussed in section 2.2 and thresholds in section 1.1.3.) The 9-

mer SPILTIITL (containing the substitution at its first position) was predicted to bind to 15 of the HLA alleles listed in section 5.2.1 when the results from NetMHCpan and NetCTLpan were aggregated. Other 9-mer spanning the same substitution were bound by an additional 3 HLA alleles from the list.

5.2.3 Selection of tumour-associated epitopes

MHC binding predictions (using the same tools and thresholds) were made for the following HCC TAA sequences: AFP (UniprotId: P02771), jglyp-3 (UniprotId: Q8IYG2), MAGE1 (UniprotId: P43355), MAGE3 (UniprotId: P43357), MAGE4 (UniprotId: P43358), NY-ESO1 (UniprotId: P78358), p53 (UniprotId: P04637) and survivin (UniprotId: O15392).

The core criteria for the selection of epitopes from these HCC TAAs was a combination of predicted binding strength and population coverage (based on the known HLA allele frequencies). However, two distinct pools of peptides were constructed based on their predicted proteasomal cleavage propensities – either high or low propensities. Predictions were made using NetCTLpan with a weighting of 0.225 for proteasomal cleavage. The logic behind the "unlikely to be cleaved" pool is that such peptides may nevertheless be presented, even if in relatively low numbers, and their (predicted) low cleavage rates may reduce the possibility that T cells capable of binding to these peptides will have been removed from the repertoire by self-tolerance mechanisms.

5.2.4 Final peptide selection

The candidate peptides generated by the preceding steps were reviewed by scientists at the Institute of Hepatology and resolved into the following four peptide pools:

Pool A: TAA peptides predicted to have a high likelihood of both cleavage and binding.

Pool B: TAA peptides predicted to have a high likelihood of binding but a low likelihood of cleavage.

Pool C: TAA peptides predicted to have a high likelihood of both cleavage and binding with more than 3 HLA alleles.

Pool D: Key TSA neoepitope SPILTIITL from p53 together with an aggregation of additional neoepitopes spanning substitution Arg249Ser (see section 5.2.2).

The rationale for Pool C was to evaluate a smaller and cheaper set of peptides than those in Pool A. (10 of the 16 peptides in Pool C are identical to peptides in Pool A.) The peptides in each pool are shown in **Table 5.2** with colour key in **Table 5.1**.

Table 5.1 Colour key for the selected peptide source protein

Tumour Antigen
Alpha-fetoprotein
Glypican 3
Melanoma-associated antigen 1
Melanoma-associated antigen 3
Melanoma-associated antigen 4
Cancer/testis antigen 1B
Cellular tumour antigen p53
Survivin

Table 5.2 The final pools of TAA and TSA peptides used experimentally

Pool A (n=32)	Pool B (n=50)	Pool C (n=16)	Pool D
APAAPTPAA	AEISLADLA	EELSVMEVY	SPILTIITL
CTYSPALNK	APRMPEAAP	ERFEMFREL	MGGMNRSPILTIITL
EELSVMEVY	ARVRFFFPS	EVDPIGHLY	
ELFDSLFPV	ARVRIAYPS	FLASFVHEY	
EVDPIGHLY	ASMELKFLI	HPFLYAPTI	
FLASFVHEY	EIARRHPFL	ISYPPLHEW	
FMNKFIYEI	ELIQKLKSF	KTYQGSYGF	
FVQENYLEY	ETYVPPAFS	LEFYLAMPF	
GSDCTTIHY	FAYYPEDLF	LESEFQAAL	

Pool A (n=32)	Pool B (n=50)	Pool C (n=16)	Pool D
ISYPPLHEW	FEFVGEFFT	MPFATPMEA	
KLKSFISFY	FLIFLLNFT	NRRPCFSSL	
KPTPASIPL	FLKDHRIST	RPILTIITL	
KTYQGSYGF	FPKTGLLII	RVRAMAIYK	
LEFYLAMPF	FPVIFGKAS	TTISFTCWR	
LYAPTILLW	FSDDKFIFH	YEIARRHPF	
MMVKPCGGY	FSDLWKLLP	YPSLTPQAF	
MPFATPMEA	FTVSGNILT		
MPKTGFLII	FYLAMPFAT		
NEISTFHNL	GEYYLQNAF		
NRRPCFSSL	IELMEVDPI		
RELNEALEL	IMPKTGFLI		
RRRELIQKL	KFIYEIARR		
RVRFFFPSL	KKHSSGCAF		
SQKTYQGSY	KLCAHSQQR		
TTINFTRQR	LWAARYDKI		
TTISFTCWR	MAITRKMAA		
WQYFFPVIF	MEQLLQSAS		
YFFPVIFSK	MPKAGLLII		
YPSLREAAL	NQLLRTMSM		
YPSLTPQAF	QEAASSSST		
YWREYILSL	QQEALGLVC		
YYLQNAFLV	RETFMNKFI		
	RMAEAGFIH		
	RTLHRNEYG		
	RTMSMPKGR		
	SQALAKRSC		
	SSCMGGMNR		

Pool A (n=32)	Pool B (n=50)	Pool C (n=16)	Pool D
	SVVGNWQYF		
	TAKKVRRAI		
	TKAEMLESV		
	TKAEMLGSV		
	TTIGKLCAH		
	VSARVRFFF		
	VTKAEMLER		
	VVGNWQYFF		
	VVRHAKNYT		
	VVRVNARVR		
	WPFLEGCAC		
	YEIARRHPF		
	YQCTAEISL		

5.3 Results

In preliminary laboratory work, scientists at the Institute of Hepatology tested PBMC from 5 healthy controls, 5 individuals with non-viral HCC and 5 individuals who developed HCC as a result of hepatitis B virus (HBV) infection. PBMC sub-samples were stimulated with each of the four peptide pools with or without the immune checkpoint inhibitor anti-PD-1. As a measure of anti-tumour immunity, ELISpot (enzyme-linked immunospot) assays were used to determine the frequency of anti-tumour cytokines IFN-γ (**Figure 5.1**) or Granzyme B, a serine protease secreted by both NK and $CD8^+$ T cells (**Figure 5.2**). Buffer with no peptide pool was included as an additional negative control, but elicited no detectable response.

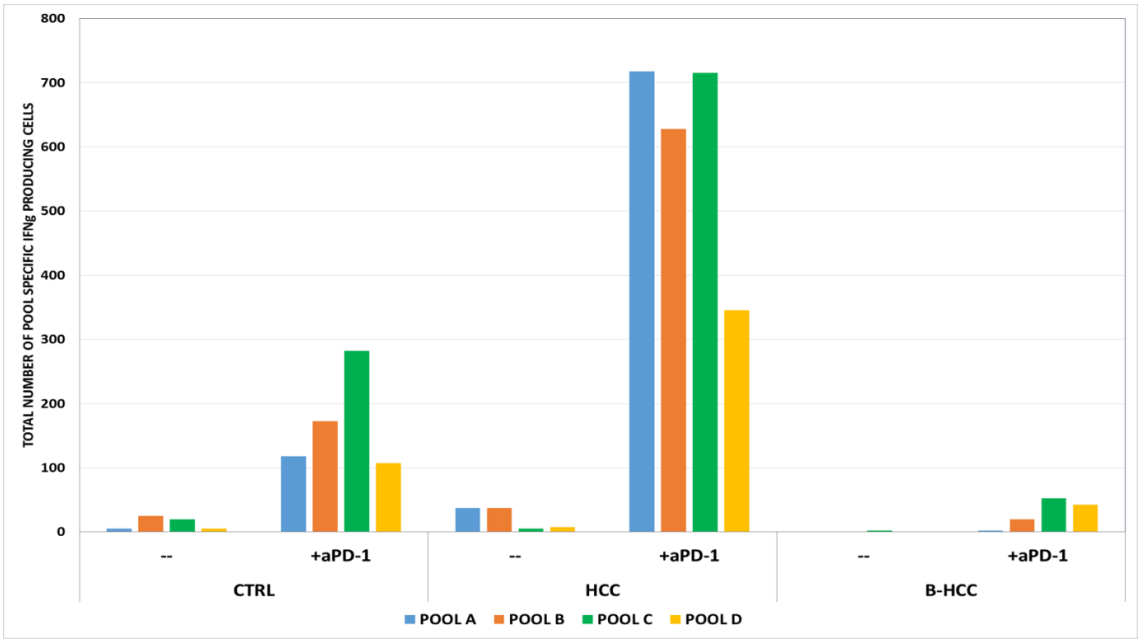


Figure 5.1 Total number of peptide pool-specific IFN-γ producing PBMCs, with and without anti-PD-1, by group. IFNg = IFN-γ, CTRL = control group, HCC = HCC without viral infection, B-HCC = HCC associated with HBV infection, +aPD-1 = in the presence of anti-PD1.

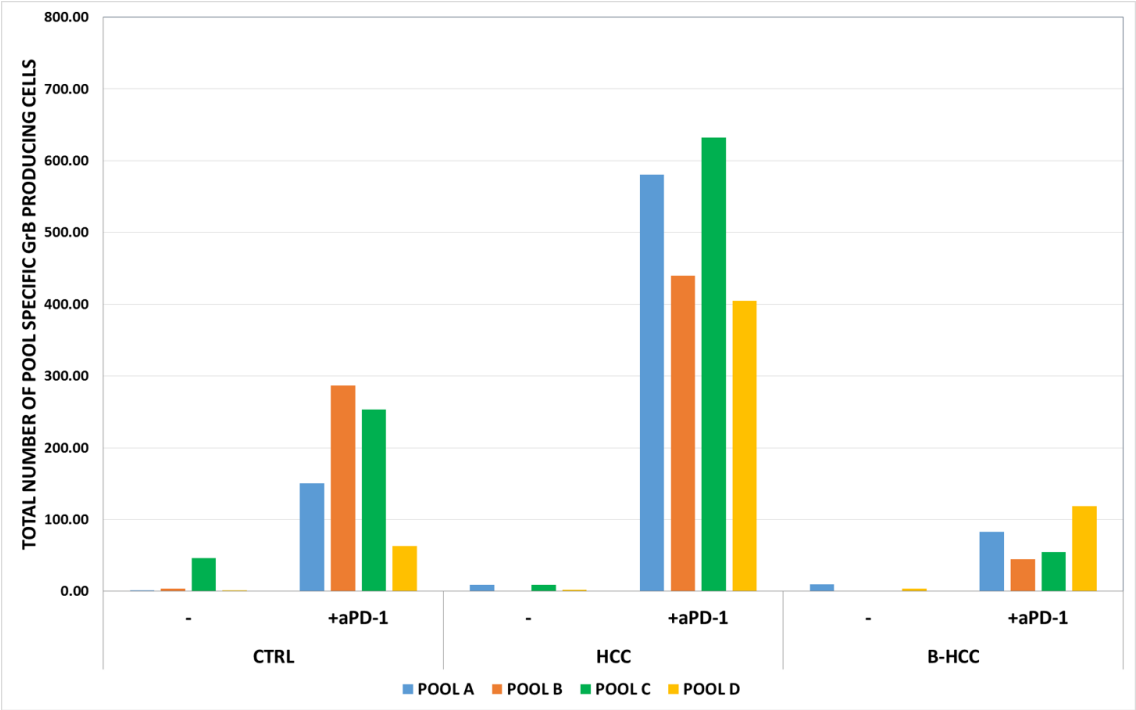


Figure 5.2 Total number of peptide pool-specific Granzyme B producing PBMCs, with and without anti-PD-1, by group. GrB = Granzyme B, other abbreviations as for Figure 5.1.

The IFN-γ (**Figure 5.1**) and Granzyme B (**Figure 5.2**) results are largely consistent and can therefore be considered together. In all cases, anti-tumour antigen immunity is low in the

absence of anti-PD-1 (with the control group Pool C response in the Granzyme B plot as perhaps a partial exception). In the presence of anti-PD-1, the highest response for all four pools is for the non-viral HCC group. The corresponding HBV-associated HCC (B-HCC) response is consistently low (though somewhat higher than nearly all the responses in the absence of anti-PD-1). The anti-PD-1control responses are intermediate between those for HCC and B-HCC, with Pool D having the lowest response.

Additional experiments of were performed using the same PBMC samples and experimental combinations with individual, whole TAAs, rather than peptides. The outcomes of these experiments are compared to the preceding peptide-based results for IFN-γ and Granzyme B in **Figure 5.3** and **Figure 5.4** respectively.

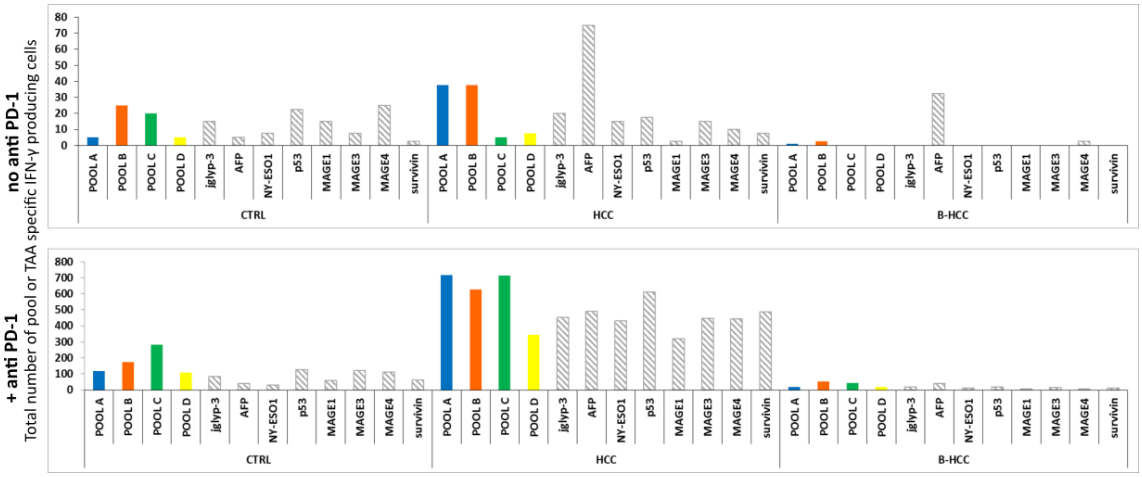


Figure 5.3 Comparison of peptide pools and individual TAAs with respect to total number of TAA-specific IFN-γ producing PBMCs.

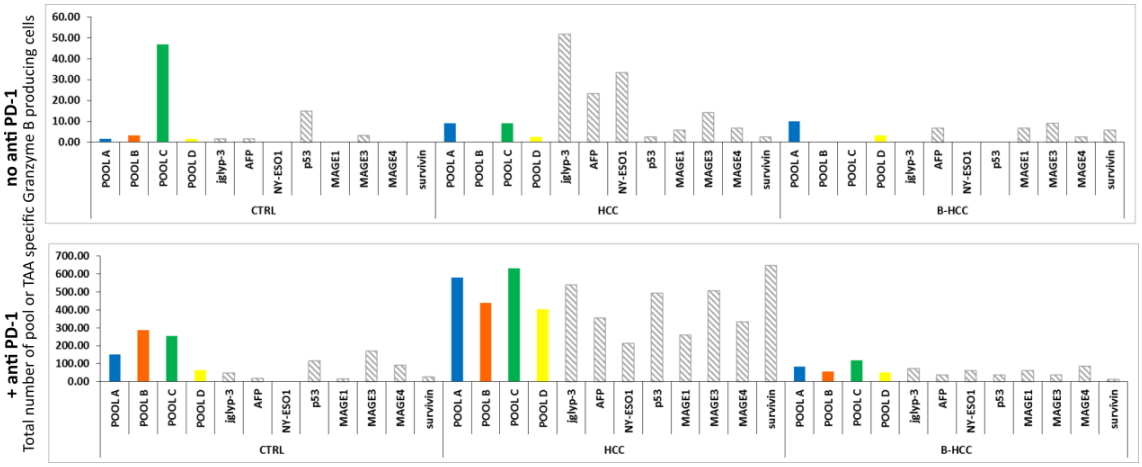


Figure 5.4 Comparison of peptide pools and individual TAAs with respect to total number of TAA-specific Granzyme B producing PBMCs.

Once again, the IFN-γ (**Figure 5.3**) and Granzyme B (**Figure 5.4**) results are largely consistent, as are the overall trends for the peptide pools in comparison to the individual TAAs.

A greater response to Pool C is seen in some of these initial experiments. This may be attributable to the fact that Pool C contains 16 peptides with high binding affinity and broad HLA coverage that may induce a strong T cell response among a large proportion of patients.

Given the relatively small number of individuals involved, further exploration of particular details risks over-interpretation.

5.4 Discussion

Two key aspects of these preliminary results are (reassuringly) in line with expectations. The restorative impact of anti-PD-1 is in line with its known efficacy (discussed in section 5.1.3); and the low immune responses detected for individuals with HBV-associated HCC are consistent with HBV's well documented suppressive impact on the host immune system, causing dysfunction in both the innate and adaptive immune responses (Li et al., 2019). Differences in the responses associated with viral versus non-viral HCC are unsurprising given the known differences between these two forms of HCC in terms of their associated protumourigenic mechanisms and impact on host T cell immunity (for a recent review, see Song &Ma, 2020).

The performance of all peptide pools can be regarded as promising, particularly with respect to the "concise" (and therefore cheap) Pool C, which suggests that optimising the choice of peptides may prove to be an effective strategy.

However, there are two key limitations with this preliminary study. Firstly, the cohort size is too small for firm conclusions to be drawn. Secondly, the results are aggregated within the groups, which means potentially vital information about individuals is unavailable. The next stage of this research will be to expand the sample cohort so that it includes a broader mixture of HLA subtypes, and to test the efficacy of the peptide pools in animal models of HCC and in *ex vivo* models such as precision cut tumour slices.

There are several key questions that need to be addressed as the study progresses. What proportion of HCC individuals are identifiable by each of the peptide pool, and to what extent is there an overlap in the individuals identified by each pool (i.e., do different individuals respond to different pools)? Are individuals with the Arg249Ser substitution consistently identified by Pool D? Although relatively modest, what does the response of control individuals to certain peptides signify (and can anything be done about it)? And what is the therapeutic potential of these peptides (e.g., in terms of a cancer vaccine)?

Finally, it is worth noting that the selection of the current peptide pools pre-dated the development of our proteome scanning pipeline (see section 3.2.2). In any future selection of peptide pools, proteome scanning could be used to filter out any cross-matching peptides, as these are unlikely to detect a host immune response, given the expectation that any T cells capable of binding to such a peptide will have been eliminated during the central tolerance process.

6 Conclusion

In this thesis, several computational techniques and data resources have been combined to address three contrasting applications. The main achievements have been:

- The development and efficient implementation of a new technique called proteome scanning that can be used, in conjunction with MHC binding predictions, to predict whether (given assumptions about the efficacy of self-tolerance mechanisms that are reasonable in most cases) a given T cell epitope is likely to be "visible" to an individual's immune system.

- Demonstrating, via two contrasting applications – inhibitor risk prediction in the context of missense mutation haemophilia A and the prediction of transplant rejection risk – that proteome scanning has broad applicability.

- Combining multiple computational techniques and data resources to make predictions that are insightful from a biological perspective and potentially impactful in three areas of biomedical importance.

The latter point deserves further elaboration. Firstly, the large-scale analysis of known haemophilia-causing Factor VIII missense mutations in combination with data about inhibitor formation in individuals with missense mutations haemophilia A has provided compelling evidence that knowledge of these individuals' HLA types is potentially important if we are to understand their inhibitor risk. Ultimately such knowledge has the potential to inform decisions about patient treatment (for example, pre-emptive immune tolerance induction might be an option for individuals with a high risk of developing inhibitors). This research also provides insights into the underlying biology: as claimed in the associated paper, "it closes part of the gap between predicted/potential inhibitor risk and observed inhibitor rates" (Hart et al., 2019). In other words, we now have a plausible explanation why many more individuals with missense mutations haemophilia A do not develop inhibitors — they are "proteome protected".

Secondly, the application of a modified proteome scanning approach to transplant rejection risk prediction was limited by available data, but nevertheless provided a compelling explanation why certain HLA residue mismatches between donor and recipient are likely to be benign and others associated with rejection risk. If this

hypothesis is confirmed by subsequent analyses with larger datasets, there is potential for proteome scanning to improve the way transplant patients are matched to potential donors.

Thirdly, the selection of HCC tumour antigen peptide pools shows early promise in terms of the detection of anti-tumour immunity, although a lot of additional laboratory work is needed before the diagnostic and/or therapeutic potential of these peptides (or a subset of them) becomes clear.

In the first two of these applications, our novel proteome scanning approach made an important contribution in terms of refining the predicted levels of risk that an adverse T cell response will develop. In the Factor VIII missense mutations case, taking a conservative threshold of IC50 < 1000 nmol/L, proteome scanning reduced the number of HLA/missense mutation combinations predicted to confer a risk of inhibitor formation from around a half to less than a third (see section 3.3.2). When predicting the risk of transplant rejection, the application of proteome scanning to the published case study (involving a HLA-B*44:03 donor versus HLA-B*44:02 recipient) reduced the number of risk-associated mismatches for a set of 25 common HLA class II alleles from 20 (80%) to only 4 (16%) (see **Figure 4.2**).

Before considering potential future work that could build on the research presented in this thesis, it is worth acknowledging some of its limitations and the challenges associated with addressing them. The boundary between self/non-self plays a pivotal role in this research, but certain factors that help to shape that boundary within an individual are difficult to model given current knowledge and available data, notably the role of the microbiome in the development of immune tolerance (see, for example, Catrina et al., 2016). Our approach to estimating risk relies on peptide-MHC binding thresholds, whereas a more effective model might consider the length of time an epitope is resident in the MHC binding groove, whether it is competing for groove occupancy with other epitopes, and the relative abundance of competing epitopes.

Notwithstanding the scale of the preceding challenges, there are several ways in which the core methodology could be improved. As noted at the end of Chapter 3, no account is currently taken of different epitope lengths, non-canonical MHC anchoring positions, and different TCR binding modes. The first may be addressed trivially by generating more

binding predictions for different peptide lengths. Some non-canonical MHC anchoring positions are document in the literature, are discernible via an examination of their crystal structures, or might be inferred from peptide-MHC binding motifs (see, for example, MHC Motif Viewer
at https://services.healthtech.dtu.dk/services/MHCMotifViewer/Home.html). Different TCR binding modes imply that certain TCRs will be in contact with fewer epitope sidechains; this is easy to model but deciding how to utilise those models in terms of their contribution to a given prediction is unclear.

However, perhaps the most important area of improvement in the context of potential clinical applications would be to provide explicit and accurate estimates of the accuracy of the predictions for a single individual (e.g., potential transplant recipient). In this context, the likely accuracy of MHC binding predictors for different HLA alleles (based on the independent evaluations described in section 2.2), are unlikely to be sufficient.

Finally, it is also worth pointing out that there are additional application areas where these approaches could be applied. In addition to protein therapeutics other than Factor VIII, there are potential applications related to infection. In particular, proteome scanning may be useful in the detection of molecular mimicry, i.e., similarity between a foreign antigen and self-antigen that is a possible causative mechanism for autoimmune disease (Cusick et al., 2012).

In all of these contexts, the availability of sufficient and appropriate data is vital for computational method development and evaluation. But this is not simply about the quantity of data; much of the current data vital to our research is susceptible to biases (e.g. T cell repertoire data comes disproportionately from individuals who come from nations that play a dominant role in undertaking scientific research). Although the desirability of removing such biases is widely recognised, a case can now be made for something more radical, given recent scientific breakthroughs involving computational tools such as AlphaFold 2 (Jumper et al., 2021): namely that computational scientists should increasingly be involved in setting the experimental agenda of wet lab scientists, so that experiments are designed to improve or validate the prediction of computational tools.

7 References

Adams, E. J., Gu, S., & Luoma, A. M. (2015). Human gamma delta T cells: Evolution and ligand recognition. In *Cellular Immunology* (Vol. 296, Issue 1, pp. 31–40). Academic Press Inc. https://doi.org/10.1016/j.cellimm.2015.04.008

Amiel, J. (1967). Study of leucocyte phenotypes in Hodgkins' disease. *Histocompatibility Testing 1967*, 79–81. https://cir.nii.ac.jp/crid/1571135649704216064.bib?lang=en

Andreatta, M., & Nielsen, M. (2016). Gapped sequence alignment using artificial neural networks: Application to the MHC class I system. *Bioinformatics*, *32*(4), 511–517. https://doi.org/10.1093/bioinformatics/btv639

Andreatta, M., Trolle, T., Yan, Z., Greenbaum, J. A., Peters, B., & Nielsen, M. (2018). An automated benchmarking platform for MHC class II binding prediction methods. *Bioinformatics*, *34*(9), 1522–1528. https://doi.org/10.1093/bioinformatics/btx820

Bagaev, D. v., Vroomans, R. M. A., Samir, J., Stervbo, U., Rius, C., Dolton, G., Greenshields-Watson, A., Attaf, M., Egorov, E. S., Zvyagin, I. v., Babel, N., Cole, D. K., Godkin, A. J., Sewell, A. K., Kesmir, C., Chudakov, D. M., Luciani, F., & Shugay, M. (2020). VDJdb in 2019: Database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Research*, *48*(D1), D1057–D1062. https://doi.org/10.1093/nar/gkz874

Baker, M. P., Reynolds, H. M., Lumicisi, B., & Bryson, C. J. (2010). Immunogenicity of protein therapeutics: The key causes, consequences and challenges. In *Self/Nonself - Immune Recognition and Signaling* (Vol. 1, Issue 4, pp. 314–322). https://doi.org/10.4161/self.1.4.13904

Barbosa, C. R. R., Barton, J., Shepherd, A. J., & Mishto, M. (2021). Mechanistic diversity in MHC class I antigen recognition. In *Biochemical Journal* (Vol. 478, Issue 24, pp. 4187–4202). Portland Press Ltd. https://doi.org/10.1042/BCJ20200910

Bartholdy, C., Reedtz-Runge, S. L., Wang, J., Hjerrild Zeuthen, L., Gruhler, A., Gudme, C. N., & Lamberth, K. (2018). In silico and in vitro immunogenicity assessment of B-domain-modified recombinant factor VIII molecules. *Haemophilia, 24*(5), e354–e362. https://doi.org/10.1111/hae.13555

Basler, M., Kirk, C. J., & Groettrup, M. (2013). The immunoproteasome in antigen processing and other immunological functions. *Current Opinion in Immunology*, *25*(1), 74–80. https://doi.org/https://doi.org/10.1016/j.coi.2012.11.004

Beringer, D. X., Kleijwegt, F. S., Wiede, F., van der Slik, A. R., Loh, K. L., Petersen, J., Dudek, N. L., Duinkerken, G., Laban, S., Joosten, A., Vivian, J. P., Chen, Z., Uldrich, A. P., Godfrey, D. I., McCluskey, J., Price, D. A., Radford, K. J., Purcell, A. W., Nikolic, T., ... Rossjohn, J. (2015). T cell receptor reversed polarity recognition of a self-antigen major histocompatibility complex. *Nature Immunology*, *16*(11), 1153–1161. https://doi.org/10.1038/ni.3271

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

Breous, E., & Thimme, R. (2011). Potential of immunotherapy for hepatocellular carcinoma. *Journal of Hepatology*, *54*(4), 830–834. https://doi.org/10.1016/j.jhep.2010.10.013

Bright, R. K., Bright, J. D., & Byrne, J. A. (2014). Overexpressed oncogenic tumor-self antigens. *Human Vaccines and Immunotherapeutics*, *10*(11), 3297–3305. https://doi.org/10.4161/hv.29475

Broere, F., & van Eden, W. (2019). T Cell Subsets and T Cell-Mediated Immunity. In M. J. Parnham, F. P. Nijkamp, & A. G. Rossi (Eds.), *Nijkamp and Parnham's Principles of Immunopharmacology* (pp. 23–35). Springer International Publishing. https://doi.org/10.1007/978-3-030-10811-3_3

Burrows, S. R., Rossjohn, J., & McCluskey, J. (2006). Have we cut ourselves too short in mapping CTL epitopes? *Trends in Immunology*, *27*(1), 11–16. https://doi.org/10.1016/j.it.2005.11.001

Catrina, A. I., Deane, K. D., & Scher, J. U. (2016). Gene, environment, microbiome and mucosal immune tolerance in rheumatoid arthritis. *Rheumatology (United Kingdom)*, *55*(3), 391–402. https://doi.org/10.1093/rheumatology/keu469

Chaves, F. A., & Sant, A. J. (2007). Measurement of Peptide Dissociation from MHC Class II Molecules. *Current Protocols in Immunology*, *77*(1), 18.14.1-18.14.11. https://doi.org/https://doi.org/10.1002/0471142735.im1814s77

Chen, W., & McCluskey, J. (2006). Immunodominance and Immunodomination: Critical Factors in Developing Effective CD8+ T-Cell–Based Cancer Vaccines. In *Advances in Cancer Research* (Vol. 95, pp. 203–247). Academic Press. https://doi.org/https://doi.org/10.1016/S0065-230X(06)95006-4

Chicz, R. M., Urban, R. G., Gorga, J. C., Vignali, D. A. A., Lane, W. S., & Strominger, J. L. (1993). Specificity and promiscuity among naturally processed peptides bound to HLA-DR alleles. In *Journal of Experimental Medicine* (Vol. 178, Issue 1). https://doi.org/10.1084/jem.178.1.27

Cloosen, S., Arnold, J., Thio, M., Bos, G. M. J., Kyewski, B., & Germeraad, W. T. V. (2007). Expression of tumor-associated differentiation antigens, MUC1 glycoforms and CEA, in human thymic epithelial cells: Implications for self-tolerance and tumor therapy. *Cancer Research*, *67*(8), 3919–3926. https://doi.org/10.1158/0008-5472.CAN-06-2112

Collins, A. M., Yaari, G., Shepherd, A. J., Lees, W., & Watson, C. T. (2020). Germline immunoglobulin genes: Disease susceptibility genes hidden in plain sight? *Current Opinion in Systems Biology*, *24*, 100–108. https://doi.org/https://doi.org/10.1016/j.coisb.2020.10.011

Corrie, B. D., Marthandan, N., Zimonja, B., Jaglale, J., Zhou, Y., Barr, E., Knoetze, N., Breden, F. M. W., Christley, S., Scott, J. K., Cowell, L. G., & Breden, F. (2018). iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. In

Immunological Reviews (Vol. 284, Issue 1, pp. 24–41). Blackwell Publishing Ltd. https://doi.org/10.1111/imr.12666

Cousens, L. P., Tassone, R., Mazer, B. D., Ramachandiran, V., Scott, D. W., & de Groot, A. S. (2013). Tregitope update: Mechanism of action parallels IVIg. In *Autoimmunity Reviews* (Vol. 12, Issue 3, pp. 436–443). https://doi.org/10.1016/j.autrev.2012.08.017

Craig, A. J., von Felden, J., Garcia-Lezana, T., Sarcognato, S., & Villanueva, A. (2020). Tumour evolution in hepatocellular carcinoma. *Nature Reviews Gastroenterology & Hepatology*, *17*(3), 139–152. https://doi.org/10.1038/s41575-019-0229-4

Cresswell, P., Bangia, N., Dick, T., & Diedrich, G. (1999). The nature of the MHC class I peptide loading complex. *Immunological Reviews*, *172*(1), 21–28. https://doi.org/https://doi.org/10.1111/j.1600-065X.1999.tb01353.x

Crispe, I. N. (2011). Liver antigen-presenting cells. *Journal of Hepatology*, *54*(2), 357–365. https://doi.org/10.1016/j.jhep.2010.10.005

Cusick, M. F., Libbey, J. E., & Fujinami, R. S. (2012). Molecular mimicry as a mechanism of autoimmune disease. *Clinical Reviews in Allergy and Immunology, 42*(1), 102–111. https://doi.org/10.1007/s12016-011-8294-7

Dandine-Roulland, C., Laurent, R., Dall'Ara, I., Toupance, B., & Chaix, R. (2019). Genomic evidence for MHC disassortative mating in humans. *Proceedings. Biological Sciences*, *286*(1899), 20182664. https://doi.org/10.1098/rspb.2018.2664

Davidson, C. J., Hirt, R. P., Lal, K., Snell, P., Elgar, G., Tuddenham, E. G. D., & McVey, J. H. (2003). Molecular evolution of the vertebrate blood coagulation network. *Thrombosis and Haemostasis*, *89*(3), 420–428.

Davis, M. M., & Boyd, S. D. (2019). Recent progress in the analysis of αβ T cell and B cell receptor repertoires. In *Current Opinion in Immunology* (Vol. 59, pp. 109–114). Elsevier Ltd. https://doi.org/10.1016/j.coi.2019.05.012

de Groot, A. S., Moise, L., McMurry, J. A., Wambre, E., van Overtvelt, L., Moingeon, P., Scott, D. W., & Martin, W. (2008). Activation of natural regulatory T cells by IgG Fc-derived peptide "Tregitopes." *Blood*, *112*(8), 3303–3311. https://doi.org/10.1182/blood-2008-02-138073

DeMaria, P. J., & Bilusic, M. (2019). Cancer Vaccines. *Hematology/Oncology Clinics of North America*, *33*(2), 199–214. https://doi.org/https://doi.org/10.1016/j.hoc.2018.12.001

Deng, L., Langley, R. J., Brown, P. H., Xu, G., Teng, L., Wang, Q., Gonzales, M. I., Callender, G. G., Nishimura, M. I., Topalian, S. L., & Mariuzza, R. A. (2007). Structural basis for the recognition of mutant self by a tumor-specific, MHC class II–restricted T cell receptor. *Nature Immunology*, *8*, 398. https://doi.org/10.1038/ni1447

Derbinski, J., & Kyewski, B. (2010). How thymic antigen presenting cells sample the body ' s. *Current Opinion in Immunology*, *22*(5), 592–600. https://doi.org/10.1016/j.coi.2010.08.003

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. http://arxiv.org/abs/1810.04805

Dorak, M. T., Lawson, T., Machulla, H. K. G., Mills, K. I., & Burnett, A. K. (2002). Increased heterozygosity for MHC class II lineages in newborn males. *Genes and Immunity*, *3*(5), 263–269. https://doi.org/10.1038/sj.gene.6363862

D'Souza, M. P., Adams, E., Altman, J. D., Birnbaum, M. E., Boggiano, C., Casorati, G., Chien, Y. H., Conley, A., Eckle, S. B. G., Früh, K., Gondré-Lewis, T., Hassan, N., Huang, H., Jayashankar, L., Kasmar, A. G., Kunwar, N., Lavelle, J., Lewinsohn, D. M., Moody, B., ... Yewdell, J. W. (2019). Casting a wider net: Immunosurveillance by nonclassical MHC molecules. In *PLoS Pathogens* (Vol. 15, Issue 2). Public Library of Science. https://doi.org/10.1371/journal.ppat.1007567

Duquesnoy, R. J. (2002). HLAMatchmaker: a molecularly based algorithm for histocompatibility determination. I. Description of the algorithm. *Human Immunology*, *63*(5), 339–352. https://doi.org/https://doi.org/10.1016/S0198-8859(02)00382-8

Duquesnoy, R. J. (2006). A Structurally Based Approach to Determine HLA Compatibility at the Humoral Immune Level. In *Human Immunology* (Vol. 67, Issue 11). https://doi.org/10.1016/j.humimm.2006.08.001

Eckhardt, C. L., van Velzen, A. S., Peters, M., Astermark, J., Brons, P. P., Castaman, G., Cnossen, M. H., Dors, N., Escuriola-Ettingshausen, C., Hamulyak, K., Hart, D. P., Hay, C. R. M., Haya, S., van Heerde, W. L., Hermans, C., Holmstr??m, M., Jimenez-Yuste, V., Keenan, R. D., Klamroth, R., ... Fijnvandraat, K. (2013). Factor VIII gene (F8) mutation and risk of inhibitor development in nonsevere hemophilia a. *Blood*. https://doi.org/10.1182/blood-2013-02-483263

Ehrenmann, F., Kaas, Q., & Lefranc, M. P. (2009). IMGT/3dstructure-DB and IMGT/domaingapalign: A database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MHcSF. *Nucleic Acids Research*, *38*(SUPPL.1). https://doi.org/10.1093/nar/gkp946

El-Khoueiry, A. B., Sangro, B., Yau, T., Crocenzi, T. S., Kudo, M., Hsu, C., Kim, T. Y., Choo, S. P., Trojan, J., Welling, T. H., Meyer, T., Kang, Y. K., Yeo, W., Chopra, A., Anderson, J., dela Cruz, C., Lang, L., Neely, J., Tang, H., ... Melero, I. (2017). Nivolumab in patients with advanced hepatocellular carcinoma (CheckMate 040): an open-label, non-comparative, phase 1/2 dose escalation and expansion trial. *The Lancet*, *389*(10088), 2492–2502. https://doi.org/10.1016/S0140-6736(17)31046-2

ElTanbouly, M. A., & Noelle, R. J. (2021). Rethinking peripheral T cell tolerance: checkpoints across a T cell's journey. *Nature Reviews Immunology*, *21*(4), 257–267. https://doi.org/10.1038/s41577-020-00454-2

Embgenbroich, M., & Burgdorf, S. (2018). Current concepts of antigen cross-presentation. In *Frontiers in Immunology* (Vol. 9, Issue JUL). Frontiers Media S.A. https://doi.org/10.3389/fimmu.2018.01643

Engel, P., Boumsell, L., Balderas, R., Bensussan, A., Gattei, V., Horejsi, V., Jin, B.-Q., Malavasi, F., Mortari, F., Schwartz-Albiez, R., Stockinger, H., van Zelm, M. C., Zola, H., & Clark, G. (2015). CD Nomenclature

2015: Human Leukocyte Differentiation Antigen Workshops as a Driving Force in Immunology. *The Journal of Immunology*, *195*(10), 4555–4563. https://doi.org/10.4049/jimmunol.1502033

Erlich, H. (2012). HLA DNA typing: Past, present, and future. *Tissue Antigens*, *80*(1), 1–11. https://doi.org/10.1111/j.1399-0039.2012.01881.x

Esfahani, K., Roudaia, L., Buhlaiga, N., Rincon, S. V. del, & Papneja, N. (2020). *A review of cancer immunotherapy : from the past , to the present , to the future*. *27*(April), 87–97.

Evnouchidou, I., Momburg, F., Papakyriakou, A., Chroni, A., Leondiadis, L., Chang, S. C., Goldberg, A. L., & Stratikos, E. (2008). The internal sequence of the peptide-substrate determines its N-Terminus trimming by ERAP1. *PLoS ONE*, *3*(11). https://doi.org/10.1371/journal.pone.0003658

Flecken, T., Schmidt, N., Hild, S., Gostick, E., Drognitz, O., Zeiser, R., Schemmer, P., Bruns, H., Eiermann, T., Price, D. A., Blum, H. E., Neumann-Haefelin, C., & Thimme, R. (2014). Immunodominance and functional alterations of tumor-associated antigen-specific CD8+ T-cell responses in hepatocellular carcinoma. *Hepatology*, *59*(4), 1415–1426. https://doi.org/10.1002/hep.26731

Fleischhauer, K., Kernan, N., O'Reilly, R., Dupont, B., & Yang, S. Y. (1990). Bone marrow-allograft rejection by T lymphocytes recognizing a single amino acid difference in HLA-B44. *The New England Journal of Medicine*.

Forabosco, P., Bouzigon, E., Ng, M. Y., Hermanowski, J., Fisher, S. A., Criswell, L. A., & Lewis, C. M. (2009). Meta-analysis of genome-wide linkage studies across autoimmune diseases. *European Journal of Human Genetics*, *17*(2), 236–243. https://doi.org/10.1038/ejhg.2008.163

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C. G., Ward, S., Dawson, E., Ponting, L., Stefancsik, R., Harsha, B., YinKok, C., Jia, M., Jubb, H., Sondka, Z., Thompson, S., De, T., & Campbell, P. J. (2017). COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Research*, *45*(D1), D777–D783. https://doi.org/10.1093/nar/gkw1121

Forner, A., Reig, M., & Bruix, J. (2018). Hepatocellular carcinoma. *The Lancet*, *391*(10127), 1301–1314. https://doi.org/10.1016/S0140-6736(18)30010-2

Franchini, M., & Mannucci, P. M. (2011). Inhibitors of propagation of coagulation (factors VIII, IX and XI): A review of current therapeutic practice. *British Journal of Clinical Pharmacology*, *72*(4), 553–562. https://doi.org/10.1111/j.1365-2125.2010.03899.x

Gardner, A., & Ruffell, B. (2016). Dendritic Cells and Cancer Immunity. In *Trends in Immunology* (Vol. 37, Issue 12, pp. 855–865). Elsevier Ltd. https://doi.org/10.1016/j.it.2016.09.006

Gielis, S., Moris, P., Bittremieux, W., de Neuter, N., Ogunjimi, B., Laukens, K., & Meysman, P. (2019). Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor Sequence Repertoires. *Frontiers in Immunology*, *10*. https://doi.org/10.3389/fimmu.2019.02820

Gonzalez-Galarza, F. F., McCabe, A., Santos, E. J. M. dos, Jones, J., Takeshita, L., Ortega-Rivera, N. D., Cid-Pavon, G. M. D., Ramsbottom, K., Ghattaoraya, G., Alfirevic, A., Middleton, D., & Jones, A. R. (2020).

Allele frequency net database (AFND) 2020 update: Gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Research*, *48*(D1), D783–D788. https://doi.org/10.1093/nar/gkz1029

Gouw, S. C., van den Berg, H. M., Oldenburg, J., Astermark, J., de Groot, P. G., Margaglione, M., Thompson, A. R., van Heerde, W., Boekhorst, J., Miller, C. H., le Cessie, S., & van der Bom, J. G. (2012). F8 gene mutation type and inhibitor development in patients with severe hemophilia A: Systematic review and meta-analysis. *Blood*, *119*(12), 2922–2934. https://doi.org/10.1182/blood-2011-09-379453

Han, Y., & Lee, A. (2021). Predicting SARS-CoV-2 epitope-specific TCR recognition using pre-trained protein embeddings. *BioRxiv*, 2021.11.17.468929. https://doi.org/10.1101/2021.11.17.468929

Harding, J. J., el Dika, I., & Abou-Alfa, G. K. (2016). Immunotherapy in hepatocellular carcinoma: Primed to make a difference? *Cancer*, *122*(3), 367–377. https://doi.org/10.1002/cncr.29769

Harndahl, M., Rasmussen, M., Roder, G., & Buus, S. (2011). Real-time, high-throughput measurements of peptide-MHC-I dissociation using a scintillation proximity assay. *Journal of Immunological Methods*, *374*(1–2), 5–12. https://doi.org/10.1016/j.jim.2010.10.012

Harndahl, M., Rasmussen, M., Roder, G., Dalgaard Pedersen, I., Sørensen, M., Nielsen, M., & Buus, S. (2012). Peptide-MHC class I stability is a better predictor than peptide affinity of CTL immunogenicity. *European Journal of Immunology*, *42*(6), 1405–1416. https://doi.org/https://doi.org/10.1002/eji.201141774

Hart, D. P., Uzun, N., Skelton, S., Kakoschke, A., Househam, J., Moss, D. S., & Shepherd, A. J. (2019). Factor VIII cross-matches to the human proteome reduce the predicted inhibitor risk in missense mutation hemophilia a. *Haematologica*, *104*(3), 599–608. https://doi.org/10.3324/haematol.2018.195669

Hasegawa, H., & Matsumoto, T. (2018). Mechanisms of tolerance induction by dendritic cells in vivo. In *Frontiers in Immunology* (Vol. 9, Issue FEB). Frontiers Media S.A. https://doi.org/10.3389/fimmu.2018.00350

Hay, C. R. M., Ludlam, C. A., Colvin, B. T., Hill, F. G. H., Preston, F. E., Wasseem, N., Bagnall, R., Peake, I. R., Organisation, the U. K. H. C. D., of, on behalf, Berntorp, E., Bunschoten, E. P. M., Fijnvandraat, K., Kasper, C. K., White, G., & Santagostino, E. (1998). Factor VIII Inhibitors in Mild and Moderate-severity Haemophilia A. *Thromb Haemost*, *79*(04), 762–766.

Herrera, O. B., Golshayan, D., Tibbott, R., Ochoa, F. S., James, M. J., Marelli-Berg, F. M., & Lechler, R. I. (2004). A Novel Pathway of Alloantigen Presentation by Dendritic Cells. *The Journal of Immunology*, *173*(8), 4828–4837. https://doi.org/10.4049/jimmunol.173.8.4828

Hill, G. R., Betts, B. C., Tkachev, V., Kean, L. S., & Blazar, B. R. (2021). Current Concepts and Advances in Graft-Versus-Host Disease Immunology. In *Annual Review of Immunology* (Vol. 39, pp. 19–49). Annual Reviews Inc. https://doi.org/10.1146/annurev-immunol-102119-073227

Holland, C. J., Cole, D. K., & Godkin, A. (2013). Re-directing CD4+ T cell responses with the flanking residues of MHC class II-bound peptides: The core is not enough. *Frontiers in Immunology*, *4*(JUL), 1–9. https://doi.org/10.3389/fimmu.2013.00172

Hoof, I., Peters, B., Sidney, J., Pedersen, L. E., Sette, A., Lund, O., Buus, S., & Nielsen, M. (2009). NetMHCpan, a method for MHC class i binding prediction beyond humans. *Immunogenetics*, *61*(1), 1–13. https://doi.org/10.1007/s00251-008-0341-z

Hsing, L. C., & Rudensky, A. Y. (2005). The lysosomal cysteine proteases in MHC class II antigen presentation. *Immunological Reviews*, *207*(1), 229–241. https://doi.org/https://doi.org/10.1111/j.0105-2896.2005.00310.x

Hui, E., Cheung, J., Zhu, J., Su, X., Taylor, M. J., Wallweber, H. A., Sasmal, D. K., Huang, J., Kim, J. M., Mellman, I., & Vale, R. D. (2017). T cell costimulatory receptor CD28 is a primary target for PD-1-mediated inhibition. *Science*, *355*(6332), 1428–1433. https://doi.org/10.1126/science.aaf1292

Ikegami, K., Mukai, M., Tsuchida, J. I., Heier, R. L., MacGregor, G. R., & Setou, M. (2006). TTLL7 is a mammalian β-tubulin polyglutamylase required for growth of MAP2-positive neurites. *Journal of Biological Chemistry*, 281(41), 30707–30716. https://doi.org/10.1074/jbc.M603984200

Jacquemin, M., Vantomme, V., Buhot, C., Lavend'homme, R., Burny, W., Demotte, N., Chaux, P., Peerlinck, K., Vermylen, J., Maillere, B., van der Bruggen, P., & Saint-Remy, J. M. (2003). CD4+ T-cell clones specific for wild-type factor VIII: A molecular mechanism responsible for a higher incidence of inhibitor formation in mild/moderate hemophilia A. *Blood*. https://doi.org/10.1182/blood-2002-05-1369

Janeway, C.A., Travers, P., Walport, M. and Shlomchik, M.J. (2001) Immunobiology: The Immune System in Health and Disease, 5th edition. Garland Science, New York.

Jensen, K. K., Andreatta, M., Marcatili, P., Buus, S., Greenbaum, J. A., Yan, Z., Sette, A., Peters, B., & Nielsen, M. (2018). Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology, 154*(3), 394–406. https://doi.org/10.1111/imm.12889

Johnsen, J. M., Fletcher, S. N., Huston, H., Roberge, S., Martin, B. K., Kircher, M., Josephson, N. C., Shendure, J., Ruuska, S., Koerper, M. A., Morales, J., Pierce, G. F., Aschman, D. J., & Konkle, B. A. (2017). Novel approach to genetic analysis and results in 3000 hemophilia patients enrolled in the My Life, Our Future initiative. *Blood Advances, 1*(13), 824–834. https://doi.org/10.1182/bloodadvances.2016002923

Josephs, T. M., Grant, E. J., & Gras, S. (2017). Molecular challenges imposed by MHC-I restricted long epitopes on T cell immunity. *Biological Chemistry*, *398*(9), 1027–1036. https://doi.org/10.1515/hsz-2016-0305

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kersh, G. J., Miley, M. J., Nelson, C. A., Grakoui, A., Horvath, S., Donermeyer, D. L., Kappler, J., Allen, P. M., Fremont, D. H., Kersh, G. J., Miley, M. J., Nelson, C. A., Grakoui, A., Horvath, S., Donermeyer, D. L.,

Kappler, J., Allen, P. M., & Fremont, D. H. (2019). *Structural and Functional Consequences of Altering a Peptide MHC Anchor Residue*. https://doi.org/10.4049/jimmunol.166.5.3345

Keşmir, C., Nussbaum, A. K., Schild, H., Detours, V., & Brunak, S. (2002). Prediction of proteasome cleavage motifs by neural networks. *Protein Engineering*, *15*(4), 287–296. https://doi.org/10.1093/protein/15.4.287

Kisselev, A. F., Akopian, T. N., Woo, K. M., & Goldberg, A. L. (1999). The sizes of peptides generated from protein by mammalian 26 and 20 S proteasomes. Implications for understanding the degradative mechanism and antigen presentation. *Journal of Biological Chemistry*, *274*(6), 3363–3371. https://doi.org/10.1074/jbc.274.6.3363

Klein, L., Kyewski, B., Allen, P. M., & Hogquist, K. A. (2014). Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nature Reviews Immunology*, *14*(6), 377–391. https://doi.org/10.1038/nri3667

Konkle, B. A., Ebbesen, L. S., Erhardtsen, E., Bianco, R. P., Lissitchkov, T., Rusen, L., & Serban, M. A. (2007). Randomized, prospective clinical trial of recombinant factor VIIa for secondary prophylaxis in hemophilia patients with inhibitors. *Journal of Thrombosis and Haemostasis*, *5*(9), 1904–1913. https://doi.org/https://doi.org/10.1111/j.1538-7836.2007.02663.x

Kotturi, M. F., Scott, I., Wolfe, T., Peters, B., Sidney, J., Cheroutre, H., von Herrath, M. G., Buchmeier, M. J., Grey, H., & Sette, A. (2008). Naive Precursor Frequencies and MHC Binding Rather Than the Degree of Epitope Diversity Shape CD8 + T Cell Immunodominance 1. In *J Immunol* (Vol. 181, Issue 3).

Kumai, T., Kobayashi, H., Harabuchi, Y., & Celis, E. (2017). Peptide vaccines in cancer — old concept revisited. In *Current Opinion in Immunology* (Vol. 45, pp. 1–7). Elsevier Ltd. https://doi.org/10.1016/j.coi.2016.11.001

Labani-Motlagh, A., Ashja-Mahdavi, M., & Loskog, A. (2020). The Tumor Microenvironment: A Milieu Hindering and Obstructing Antitumor Immune Responses. In *Frontiers in Immunology* (Vol. 11). Frontiers Media S.A. https://doi.org/10.3389/fimmu.2020.00940

Larsen, M. v, Lundegaard, C., Lamberth, K., Buus, S., Lund, O., & Nielsen, M. (2007). Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinformatics*, *8*, 424.

Laydon, D. J., Bangham, C. R. M., & Asquith, B. (2015). Estimating T-cell repertoire diversity: Limitations of classical estimators and a new approach. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1675). https://doi.org/10.1098/rstb.2014.0291

Lazarski, C. A., Chaves, F. A., Jenks, S. A., Wu, S., Richards, K. A., Weaver, J. M., & Sant, A. J. (2005). The kinetic stability of MHC class II:Peptide complexes is a key parameter that dictates immunodominance. *Immunity*, *23*(1), 29–40. https://doi.org/10.1016/j.immuni.2005.05.009

Lee, S. J., Klein, J., Haagenson, M., Baxter-Lowe, L. A., Confer, D. L., Eapen, M., Fernandez-Vina, M., Flomenberg, N., Horowitz, M., Hurley, C. K., Noreen, H., Oudshoorn, M., Petersdorf, E., Setterholm, M., Spellman, S., Weisdorf, D., Williams, T. M., & Anasetti, C. (2007). High-resolution donor-recipient HLA

matching contributes to the success of unrelated donor marrow transplantation. *Blood*, *110*(13), 4576–4583. https://doi.org/10.1182/blood-2007-06-097386

Lefranc, M.-P., & Lefranc, G. (2001). *The T cell receptor FactsBook*.

Li, M. O., & Rudensky, A. Y. (2016). T cell receptor signalling in the control of regulatory T cell differentiation and function. *Nature Reviews Immunology*, *16*(4), 220–233. https://doi.org/10.1038/nri.2016.26

Li, T. Y., Yang, Y., Zhou, G., & Tu, Z. K. (2019). Immune suppression in chronic hepatitis B infection associated liver disease: A review. In *World Journal of Gastroenterology* (Vol. 25, Issue 27, pp. 3527–3537). Baishideng Publishing Group Co. https://doi.org/10.3748/wjg.v25.i27.3527

Liepe, J., Marino, F., Sidney, J., Jeko, A., Bunting, D. E., Sette, A., Kloetzel, P. M., Stumpf, M. P. H., Heck, A. J. R., & Mishto, M. (2016). A large fraction of HLA class I ligands are proteasome-generated spliced peptides. In *Science* (Vol. 354, Issue 6310). https://doi.org/10.1126/science.aaf4384

Lieuw, K. (2017). Many factor VIII products available in the treatment of hemophilia a: An embarrassment of riches? *Journal of Blood Medicine*, *8*, 67–73. https://doi.org/10.2147/JBM.S1037

Lin, C. M., & Gill, R. G. (2016). Direct and indirect allograft recognition: Pathways dictating graft rejection mechanisms. *Current Opinion in Organ Transplantation*, *21*(1), 139–148. https://doi.org/10.1097/MOT.0000000000000263

Lippolis, J. D., White, F. M., Marto, J. A., Luckey, C. J., Bullock, T. N. J., Shabanowitz, J., Hunt, D. F., & Engelhard, V. H. (2002). Analysis of MHC Class II Antigen Processing by Quantitation of Peptides that Constitute Nested Sets. *The Journal of Immunology*, *169*(9), 5089–5097. https://doi.org/10.4049/jimmunol.169.9.5089

Loiseau, P., Busson, M., Balere, M. L., Dormoy, A., Bignon, J. D., Gagne, K., Gebuhrer, L., Dubois, V., Jollet, I., Bois, M., Perrier, P., Masson, D., Moine, A., Absi, L., Reviron, D., Lepage, V., Tamouza, R., Toubert, A., Marry, E., ... Raffoux, C. (2007). HLA Association with Hematopoietic Stem Cell Transplantation Outcome: The Number of Mismatches at HLA-A, -B, -C, -DRB1, or -DQB1 Is Strongly Associated with Overall Survival. *Biology of Blood and Marrow Transplantation*, *13*(8), 965–974. https://doi.org/10.1016/j.bbmt.2007.04.010

Lv, H. J., Havari, E., Pinto, S., Gottumukkala, R. V. S. R. K., Cornivelli, L., Raddassi, K., Matsui, T., Rosenzweig, A., Bronson, R. T., Smith, R., Fletcher, A. L., Turley, S. J., Wucherpfennig, K., Kyewski, B., & Lipes, M. A. (2011). Impaired thymic tolerance to α-myosin directs autoimmunity to the heart in mice and humans. *Journal of Clinical Investigation*, *121*(4), 1561–1573. https://doi.org/10.1172/JCI44583

Marques, A. J., Palanimurugan, R., Mafias, A. C., Ramos, P. C., & Dohmen, R. J. (2009). Catalytic mechanism and assembly of the proteasome. *Chemical Reviews*, *109*(4), 1509–1536. https://doi.org/10.1021/cr8004857

Mayor, N. P., Hayhurst, J. D., Turner, T. R., Szydlo, R. M., Shaw, B. E., Bultitude, W. P., Sayno, J. R., Tavarozzi, F., Latham, K., Anthias, C., Robinson, J., Braund, H., Danby, R., Perry, J., Wilson, M. C., Bloor, A. J., McQuaker, I. G., MacKinnon, S., Marks, D. I., ... Marsh, S. G. E. (2019). Recipients

Receiving Better HLA-Matched Hematopoietic Cell Transplantation Grafts, Uncovered by a Novel HLA Typing Method, Have Superior Survival: A Retrospective Study. *Biology of Blood and Marrow Transplantation,* 25(3), 443–450. https://doi.org/10.1016/j.bbmt.2018.12.768

Mayor, N. P., Robinson, J., McWhinnie, A. J. M., Ranade, S., Eng, K., Midwinter, W., Bultitude, W. P., Chin, C. S., Bowman, B., Marks, P., Braund, H., Madrigal, J. A., Latham, K., & Marsh, S. G. E. (2015). HLA typing for the next generation. *PLoS ONE, 10*(5). https://doi.org/10.1371/journal.pone.0127153

Meurer, T., Crivello, P., Metzing, M., Kester, M., Megger, D. A., Chen, W., van Veelen, P. A., van Balen, P., Westendorf, A. M., Homa, G., Layer, S. E., Turki, A. T., Griffioen, M., Horn, P. A., Sitek, B., Beelen, D. W., Falkenburg, J. H. F., Arrieta-Bolaños, E., & Fleischhauer, K. (2021). Permissive HLA-DPB1 mismatches in HCT depend on immunopeptidome divergence and editing by HLA-DM. In *Blood* (Vol. 137, Issue 7). https://doi.org/10.1182/blood.2020008464

Milner, E., Gutter-Kapon, L., Bassani-Strenberg, M., Barnea, E., Beer, I., & Admon, A. (2013). The effect of proteasome inhibition on the generation of the human leukocyte antigen (HLA) peptidome. *Molecular and Cellular Proteomics*, *12*(7), 1853–1864. https://doi.org/10.1074/mcp.M112.026013

Moini, M., Schilsky, M. L., & Tichy, E. M. (2015). Review on immunosuppression in liver transplantation. In *World Journal of Hepatology* (Vol. 7, Issue 10, pp. 1355–1368). Baishideng Publishing Group Co. https://doi.org/10.4254/wjh.v7.i10.1355

Montemurro, A., Schuster, V., Povlsen, H. R., Bentzen, A. K., Jurtz, V., Chronister, W. D., Crinklaw, A., Hadrup, S. R., Winther, O., Peters, B., Jessen, L. E., & Nielsen, M. (2021). NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data. *Communications Biology*, *4*(1). https://doi.org/10.1038/s42003-021-02610-3

Mylonas, R., Beer, I., Iseli, C., Chong, C., Pak, H. S., Gfeller, D., Coukos, G., Xenarios, I., Müller, M., & Bassani-Sternberg, M. (2018). Estimating the contribution of proteasomal spliced peptides to the HLA-I ligandome. *Molecular and Cellular Proteomics*, *17*(12), 2347–2357. https://doi.org/10.1074/mcp.RA118.000877

Nakagawa, H., Mizukoshi, E., Kobayashi, E., Tamai, T., Hamana, H., Ozawa, T., Kishi, H., Kitahara, M., Yamashita, T., Arai, K., Terashima, T., Iida, N., Fushimi, K., Muraguchi, A., & Kaneko, S. (2017). Association Between High-Avidity T-Cell Receptors, Induced by α-Fetoprotein–Derived Peptides, and Anti-Tumor Effects in Patients With Hepatocellular Carcinoma. *Gastroenterology*, *152*(6), 1395-1406.e10. https://doi.org/10.1053/j.gastro.2017.02.001

Nakano, S., Eso, Y., Okada, H., Takai, A., Takahashi, K., & Seno, H. (2020). Recent advances in immunotherapy for hepatocellular carcinoma. *Cancers*, *12*(4), 9–10. https://doi.org/10.3390/cancers12040775

Neefjes, J., Jongsma, M. L. M., Paul, P., & Bakke, O. (2011). Towards a systems understanding of MHC class i and MHC class II antigen presentation. In *Nature Reviews Immunology* (Vol. 11, Issue 12, pp. 823–836). https://doi.org/10.1038/nri3084

Nielsen, M., & Lund, O. (2009). NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinformatics*, *10*, 296.

Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., Røder, G., Peters, B., Sette, A., Lund, O., & Buus, S. (2007). NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS ONE*, *2*(8), e796. https://doi.org/10.1371/journal.pone.0000796

Nielsen, M., Lundegaard, C., Lund, O., & Keşmir, C. (2005). The role of the proteasome in generating cytotoxic T-cell epitopes: Insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics*, *57*(1–2), 33–41. https://doi.org/10.1007/s00251-005-0781-7

Nielsen, M., Lundegaard, C., Worning, P., Lauemøller, S. L., Lamberth, K., Buus, S., Brunak, S., & Lund, O. (2003). Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Science : A Publication of the Protein Society*, *12*(5), 1007–1017.

O'Brien, C., Flower, D. R., & Feighery, C. (2008). Peptide length significantly influences in vitro affinity for MHC class II molecules. *Immunome Research*, *4*(1). https://doi.org/10.1186/1745-7580-4-6

O'Donoghue, A. J., Alegra Eroy-Reveles, A. A., Knudsen, G. M., Ingram, J., Zhou, M., Statnekov, J. B., Greninger, A. L., Hostetter, D. R., Qu, G., Maltby, D. A., Anderson, M. O., Derisi, J. L., McKerrow, J. H., Burlingame, A. L., & Craik, C. S. (2012). Global identification of peptidase specificity by multiplex substrate profiling. *Nature Methods*, *9*(11), 1095–1100. https://doi.org/10.1038/nmeth.2182

Omer, A., Peres, A., Rodriguez, O. L., Watson, C. T., Lees, W., Polak, P., Collins, A. M., & Yaari, G. (2022). T cell receptor beta germline variability is revealed by inference from repertoire data. *Genome Medicine*, *14*(1). https://doi.org/10.1186/s13073-021-01008-4

Paul, S., Grifoni, A., Peters, B., & Sette, A. (2020). Major Histocompatibility Complex Binding, Eluted Ligands, and Immunogenicity: Benchmark Testing and Predictions. *Frontiers in Immunology*, *10*(February), 1–10. https://doi.org/10.3389/fimmu.2019.03151

Paul, S., Lindestam Arlehamn, C. S., Scriba, T. J., Dillon, M. B. C., Oseroff, C., Hinz, D., McKinney, D. M., Carrasco Pro, S., Sidney, J., Peters, B., & Sette, A. (2015). Development and validation of a broad scheme for prediction of HLA class II restricted T cell epitopes. *Journal of Immunological Methods*, *422*, 28–34. https://doi.org/10.1016/j.jim.2015.03.022

Paul, S., Weiskopf, D., Angelo, M. A., Sidney, J., Peters, B., & Sette, A. (2013). HLA Class I Alleles Are Associated with Peptide-Binding Repertoires of Different Size, Affinity, and Immunogenicity. *The Journal of Immunology*, *191*(12), 5831–5839. https://doi.org/10.4049/jimmunol.1302101

Perreault, C., Decary, F., Brochu, S., Gyger, M., Belanger, R., & Roy, D. (1990). Minor Histocompatibility Antigens. *Blood*, *76*(7), 1269–1280. https://doi.org/https://doi.org/10.1182/blood.V76.7.1269.1269

Peters, B., Nielsen, M., & Sette, A. (2020). T Cell Epitope Predictions. *Annual Review of Immunology*, *38*(1), 123–145. https://doi.org/10.1146/annurev-immunol-082119-124838

Petrova, G., Ferrante, A., & Gorski, J. (2012). *Cross-Reactivity of T Cells and Its Role in the Immune System*.

Purcell, A. W. (2021). Is the Immunopeptidome Getting Darker?: A Commentary on the Discussion around Mishto et al., 2019. *Frontiers in Immunology*, *12*. https://doi.org/10.3389/fimmu.2021.720811

Rammensee, H., Bachmann, J., Emmerich, N. P., Bachor, O. A., & Stevanović, S. (1999). SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, *50*(3–4), 213–219.

Rapin, N., Hoof, I., Lund, O., & Nielsen, M. (2008). MHC motif viewer. *Immunogenetics*, *60*(12), 759–765. https://doi.org/10.1007/s00251-008-0330-2

Reardon, D. A., & Wen, P. Y. (2015). Unravelling tumour heterogeneity—implications for therapy. *Nature Reviews Clinical Oncology*, *12*(2), 69–70. https://doi.org/10.1038/nrclinonc.2014.223

Reynisson, B., Barra, C., Kaabinejadian, S., Hildebrand, W. H., Peters, B., Peters, B., Nielsen, M., & Nielsen, M. (2020). Improved Prediction of MHC II Antigen Presentation through Integration and Motif Deconvolution of Mass Spectrometry MHC Eluted Ligand Data. *Journal of Proteome Research*, *19*(6), 2304–2315. https://doi.org/10.1021/acs.jproteome.9b00874

Ringelhan, M., Pfister, D., O'Connor, T., Pikarsky, E., & Heikenwalder, M. (2018). The immunology of hepatocellular carcinoma review-article. *Nature Immunology*, *19*(3), 222–232. https://doi.org/10.1038/s41590-018-0044-z

Robins, H. S., Campregher, P. v., Srivastava, S. K., Wacher, A., Turtle, C. J., Kahsai, O., Riddell, S. R., Warren, E. H., & Carlson, C. S. (2009). Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells. *Blood*, *114*(19), 4099–4107. https://doi.org/10.1182/blood-2009-04-217604

Robinson, J., Barker, D. J., Georgiou, X., Cooper, M. A., Flicek, P., & Marsh, S. G. E. (2020). IPD-IMGT/HLA Database. *Nucleic Acids Research*, *48*(D1), D948–D955. https://doi.org/10.1093/nar/gkz950

Rock, K. L., York, I. A., & Goldberg, A. L. (2004). Post-proteasomal antigen processing for major histocompatibility complex class I presentation. *Nature Immunology*, *5*(7), 670–677. https://doi.org/10.1038/ni1089

Rohaan, M. W., Wilgenhof, S., & Haanen, J. B. A. G. (2019). Adoptive cellular therapies: the current landscape. *Virchows Archiv*, *474*(4), 449–461. https://doi.org/10.1007/s00428-018-2484-0

Rossjohn, J., Gras, S., Miles, J. J., Turner, S. J., Godfrey, D. I., & McCluskey, J. (2015). T Cell Antigen Receptor Recognition of Antigen-Presenting Molecules. *Annual Review of Immunology*, *33*(1), 169–200. https://doi.org/10.1146/annurev-immunol-032414-112334

Rudolph, M. G., Stanfield, R. L., & Wilson, I. a. (2006). How TCRs bind MHCs, peptides, and coreceptors. *Annual Review of Immunology*, *24*, 419–466. https://doi.org/10.1146/annurev.immunol.23.021704.115658

Sant, A. J., Chaves, F. A., Krafcik, F. R., Lazarski, C. A., Menges, P., Richards, K., & Weaver, J. M. (2007). Immunodominance in CD4 T-cell responses: implications for immune responses to influenza virus and for vaccine design. *Expert Review of Vaccines*, *6*(3), 357–368. https://doi.org/10.1586/14760584.6.3.357

Sauna, Z. E., Ameri, A., Kim, B., Yanover, C., Viel, K. R., Rajalingam, R., Cole, S. A., & Howard, T. E. (2012). Observations regarding the immunogenicity of BDD-rFVIII derived from a mechanistic personalized medicine perspective. In *Journal of thrombosis and haemostasis : JTH* (Vol. 10, Issue 9, pp. 1961–1965). https://doi.org/10.1111/j.1538-7836.2012.04830.x

Sawada, Y., Yoshikawa, T., Nobuoka, D., Shirakawa, H., Kuronuma, T., Motomura, Y., Mizuno, S., Ishii, H., Nakachi, K., Konishi, M., Nakagohri, T., Takahashi, S., Gotohda, N., Takayama, T., Yamao, K., Uesaka, K., Furuse, J., Kinoshita, T., & Nakatsura, T. (2012). Phase I trial of a glypican-3-derived peptide vaccine for advanced hepatocellular carcinoma: Immunologic evidence and potential for improving overall survival. *Clinical Cancer Research*, *18*(13), 3686–3696. https://doi.org/10.1158/1078-0432.CCR-11-3044

Sette, A., Vitiello, A., Reherman, B., Fowler, P., Nayersina, R., Kast, W. M., Melief, C. J., Oseroff, C., Yuan, L., Ruppert, J., Sidney, J., del Guercio, M. F., Southwood, S., Kubo, R. T., Chesnut, R. W., Grey, H. M., & Chisari, F. v. (1994). The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *Journal of Immunology (Baltimore, Md. : 1950)*, *153*(12), 5586–5592. http://www.ncbi.nlm.nih.gov/pubmed/7527444

Sewell, A. K. (2012). Why must T cells be cross-reactive? In *Nature Reviews Immunology* (Vol. 12, Issue 9, pp. 669–677). https://doi.org/10.1038/nri3279

Shepherd, A. J., Skelton, S., Sansom, C. E., Gomez, K., Moss, D. S., & Hart, D. P. (2015). A large-scale computational study of inhibitor risk in non-severe haemophilia A. *British Journal of Haematology*, *168*(3), 413–420. https://doi.org/10.1111/bjh.13131

Shiina, T., Hosomichi, K., Inoko, H., & Kulski, J. K. (2009). The HLA genomic loci map: Expression, interaction, diversity and disease. *Journal of Human Genetics*, *54*(1), 15–39. https://doi.org/10.1038/jhg.2008.5

Shugay, M., Bolotin, D. A., Putintseva, E. v., Pogorelyy, M. v., Mamedov, I. Z., & Chudakov, D. M. (2013). Huge overlap of individual TCR beta repertoires. In *Frontiers in Immunology* (Vol. 4, Issue DEC). https://doi.org/10.3389/fimmu.2013.00466

Song, X.-J., & Ma, C.-H. (2020). Mechanisms and immunotherapies of HBV- and NAFLD-related hepatocellular carcinoma. *Hepatoma Research, 2020*. https://doi.org/10.20517/2394-5079.2020.05

Soria-Guerra, R. E., Nieto-Gomez, R., Govea-Alonso, D. O., & Rosales-Mendoza, S. (2015). An overview of bioinformatics tools for epitope prediction: Implications on vaccine development. *Journal of Biomedical Informatics*, *53*, 405–414. https://doi.org/10.1016/j.jbi.2014.11.003

Southwood, S., Sidney, J., Kondo, A., del Guercio, M. F., Appella, E., Hoffman, S., Kubo, R. T., Chesnut, R. W., Grey, H. M., & Sette, A. (1998). Several common HLA-DR types share largely overlapping peptide binding repertoires. *Journal of Immunology (Baltimore, Md. : 1950)*, *160*(7), 3363–3373. http://www.ncbi.nlm.nih.gov/pubmed/9531296

Stranzl, T., Larsen, M. V., Lundegaard, C., & Nielsen, M. (2010). NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics*, *62*(6), 357–368.

Su, C. (2016). Survivin in survival of hepatocellular carcinoma. *Cancer Letters*, *379*(2), 184–190. https://doi.org/10.1016/j.canlet.2015.06.016

Swain, S. L., McKinstry, K. K., & Strutt, T. M. (2012). Expanding roles for CD4 + T cells in immunity to viruses. In *Nature Reviews Immunology* (Vol. 12, Issue 2, pp. 136–148). https://doi.org/10.1038/nri3152

Toes, R. E. M., Nussbaum, A. K., Degermann, S., Schirle, M., Emmerich, N. P. N., Kraft, M., Laplace, C., Zwinderman, A., Dick, T. P., Müller, J., Schönfisch, B., Schmid, C., Fehling, H.-J., Stevanovic, S., Rammensee, H. G., & Schild, H. (2001). Discrete Cleavage Motifs of Constitutive and Immunoproteasomes Revealed by Quantitative Analysis of Cleavage Products. In *J. Exp. Med. ‣ The* (Vol. 194, Issue 1). Rockefeller University Press. http://www.jem.org/cgi/content/full/194/1/1

Trolle, T., Metushi, I. G., Greenbaum, J. A., Kim, Y., Sidney, J., Lund, O., Sette, A., Peters, B., & Nielsen, M. (2015). Automated benchmarking of peptide-MHC class i binding predictions. *Bioinformatics*, *31*(13), 2174–2181. https://doi.org/10.1093/bioinformatics/btv123

Tscharke, D. C., Croft, N. P., Doherty, P. C., & la Gruta, N. L. (2015). Sizing up the key determinants of the CD8+ T cell response. In *Nature Reviews Immunology* (Vol. 15, Issue 11, pp. 705–716). Nature Publishing Group. https://doi.org/10.1038/nri3905

Valujskikh, A., Lantz, O., Celli, S., Matzinger, P., & Heeger, P. S. (2002). Cross-primed CD8+ T cells mediate graft rejection via a distinct effector pathway. *Nature Immunology*, *3*(9), 844–851. https://doi.org/10.1038/ni831

van der Burg, S. H., Visseren, M. J., Brandt, R. M., Kast, W. M., & Melief, C. J. (1996). Immunogenicity of peptides bound to MHC class I molecules depends on the MHC-peptide complex stability. *The Journal of Immunology*, *156*(9), 3308. http://www.jimmunol.org/content/156/9/3308.abstract

Vanhanen, R., Heikkilä, N., Aggarwal, K., Hamm, D., Tarkkila, H., Pätilä, T., Jokiranta, T. S., Saramäki, J., & Arstila, T. P. (2016). T cell receptor diversity in the human thymus. *Molecular Immunology*, *76*, 116–122. https://doi.org/10.1016/j.molimm.2016.07.002

Via, M., Gignoux, C., & Burchard, E. G. (2010). The 1000 Genomes Project: New opportunities for research and social challenges. In *Genome Medicine* (Vol. 2, Issue 1). https://doi.org/10.1186/gm124

Viel, K. R., Ameri, A., Abshire, T. C., Iyer, R. v., Watts, R. G., Lutcher, C., Channell, C., Cole, S. A., Fernstrom, K. M., Nakaya, S., Kasper, C. K., Thompson, A. R., Almasy, L., & Howard, T. E. (2009). Inhibitors of Factor VIII in Black Patients with Hemophilia. *New England Journal of Medicine*, *360*(16), 1618–1627. https://doi.org/10.1056/nejmoa075760

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., & Peters, B. (2019). The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research*, *47*(D1), D339–D343. https://doi.org/10.1093/nar/gky1006

Vizcaíno, J. A., Kubiniok, P., Kovalchik, K. A., Ma, Q., Duquette, J. D., Mongrain, I., Deutsch, E. W., Peters, B., Sette, A., Sirois, I., & Caron, E. (2020). The human immunopeptidome project: A roadmap to predict and treat immune diseases. In *Molecular and Cellular Proteomics* (Vol. 19, Issue 1, pp. 31–49). American Society for Biochemistry and Molecular Biology Inc. https://doi.org/10.1074/mcp.R119.001743

Vousden, K. H., & Lane, D. P. (2007). P53 in Health and Disease. *Nature Reviews Molecular Cell Biology*, *8*(4), 275–283. https://doi.org/10.1038/nrm2147

Walker, J. A., & McKenzie, A. N. J. (2018). TH2 cell development and function. *Nature Reviews Immunology*, *18*(2), 121–133. https://doi.org/10.1038/nri.2017.118

Wang, H., Xu, J., Lazarovici, P., & Zheng, W. (2017). Dysbindin-1 involvement in the etiology of schizophrenia. In *International Journal of Molecular Sciences* (Vol. 18, Issue 10). MDPI AG. https://doi.org/10.3390/ijms18102044

Wang, J. H., Skeans, M. A., & Israni, A. K. (2016). Current Status of Kidney Transplant Outcomes: Dying to Survive. *Advances in Chronic Kidney Disease*, *23 5*, 281–286.

Wang, M., Li, J., Wang, L., Chen, X., Zhang, Z., Yue, D., Ping, Y., Shi, X., Huang, L., Zhang, T., Yang, L., Zhao, Y., Ma, X., Li, D., Fan, Z., Zhao, L., Tang, Z., Zhai, W., Zhang, B., & Zhang, Y. (2015). Combined cancer testis antigens enhanced prediction accuracy for prognosis of patients with hepatocellular carcinoma. *International Journal of Clinical and Experimental Pathology*, *8*(4), 3513–3528.

Wang, P., Sidney, J., Kim, Y., Sette, A., Lund, O., Nielsen, M., & Peters, B. (2010). Peptide binding predictions for HLA DR, DP and DQ molecules. *BMC Bioinformatics*, *11*, 568. https://doi.org/10.1186/1471-2105-11-568

Wasmuth, E. v, & Lima, C. D. (2016). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, *45*(November 2016), 1–12. https://doi.org/10.1093/nar/gkw1152

Wei, S. C., Duffy, C. R., & Allison, J. P. (2018). Fundamental mechanisms of immune checkpoint blockade therapy. *Cancer Discovery*, *8*(9), 1069–1086. https://doi.org/10.1158/2159-8290.CD-18-0367

Wills, M. R., Carmichael, A. J., Mynard, K., Jin, X., Weekes, M. P., Plachter, B., & Sissons, J. G. P. (1996). The Human Cytotoxic T-Lymphocyte (CTL) Response to Cytomegalovirus Is Dominated by Structural Protein pp65: Frequency, Specificity, and T-Cell Receptor Usage of pp65-Specific CTL. In *JOURNAL OF VIROLOGY* (Vol. 70, Issue 11).

Winter, M. B., la Greca, F., Arastu-Kapur, S., Caiazza, F., Cimermancic, P., Buchholz, T. J., Anderl, J. L., Ravalin, M., Bohn, M. F., Sali, A., O, A. J., & Craik, C. S. (n.d.). *Immunoproteasome functions explained by divergence in cleavage specificity and regulation*. https://doi.org/10.7554/eLife.27364.001

Wood, K. J., & Goto, R. (2012). Mechanisms of rejection: Current perspectives. *Transplantation*, *93*(1), 1–10. https://doi.org/10.1097/TP.0b013e31823cab44

Wooldridge, L., Ekeruche-Makinde, J., van den Berg, H. A., Skowera, A., Miles, J. J., Tan, M. P., Dolton, G., Clement, M., Llewellyn-Lacey, S., Price, D. A., Peakman, M., & Sewell, A. K. (2012). A single autoimmune

T cell receptor recognizes more than a million different peptides. *Journal of Biological Chemistry*, *287*(2), 1168–1177. https://doi.org/10.1074/jbc.M111.289488

Wu, K. E., Yost, K. E., Daniel, B., Belk, J. A., Xia, Y., Egawa, T., Satpathy, A., Chang, H., & Zou, J. (2021). TCR-BERT: learning the grammar of T-cell receptors for flexible antigen- binding analyses. *BioRxiv*, 2021.11.18.469186. https://doi.org/https://doi.org/10.1101/2021.11.18.469186

Yang, X., Gao, M., Chen, G., Pierce, B. G., Lu, J., Weng, N. P., & Mariuzza, R. A. (2015). Structural basis for clonal diversity of the public T cell response to a dominant human cytomegalovirus epitope. *Journal of Biological Chemistry*, *290*(48), 29106–29119. https://doi.org/10.1074/jbc.M115.691311

Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gordon, L., Hourlier, T., Hunt, S. E., Gil, L., Garc, C., Janacek, H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., … Flicek, P. (2016). *Ensembl 2016*. *44*(December 2015), 710–716. https://doi.org/10.1093/nar/gkv1157

Yau, T., Park, J. W., Finn, R. S., Cheng, A.-L., Mathurin, P., Edeline, J., Kudo, M., Han, K.-H., Harding, J. J., Merle, P., Rosmorduc, O., Wyrwicz, L., Schott, E., Choo, S. P., Kelley, R. K., Begic, D., Chen, G., Neely, J., Anderson, J., & Sangro, B. (2019). CheckMate 459: A randomized, multi-center phase III study of nivolumab (NIVO) vs sorafenib (SOR) as first-line (1L) treatment in patients (pts) with advanced hepatocellular carcinoma (aHCC). *Annals of Oncology*, *30*(October), v874–v875. https://doi.org/10.1093/annonc/mdz394.029

Yewdell, J. W. (2006). Confronting complexity: real-world immunodominance in antiviral CD8+ T cell responses. *Immunity*, *25*(4), 533–543. https://doi.org/10.1016/j.immuni.2006.09.005

Yin, Y., Li, Y., & Mariuzza, R. A. (2012). Structural basis for self-recognition by autoimmune T-cell receptors. *Immunological Reviews*, *250*(1), 32–48. https://doi.org/https://doi.org/10.1111/imr.12002

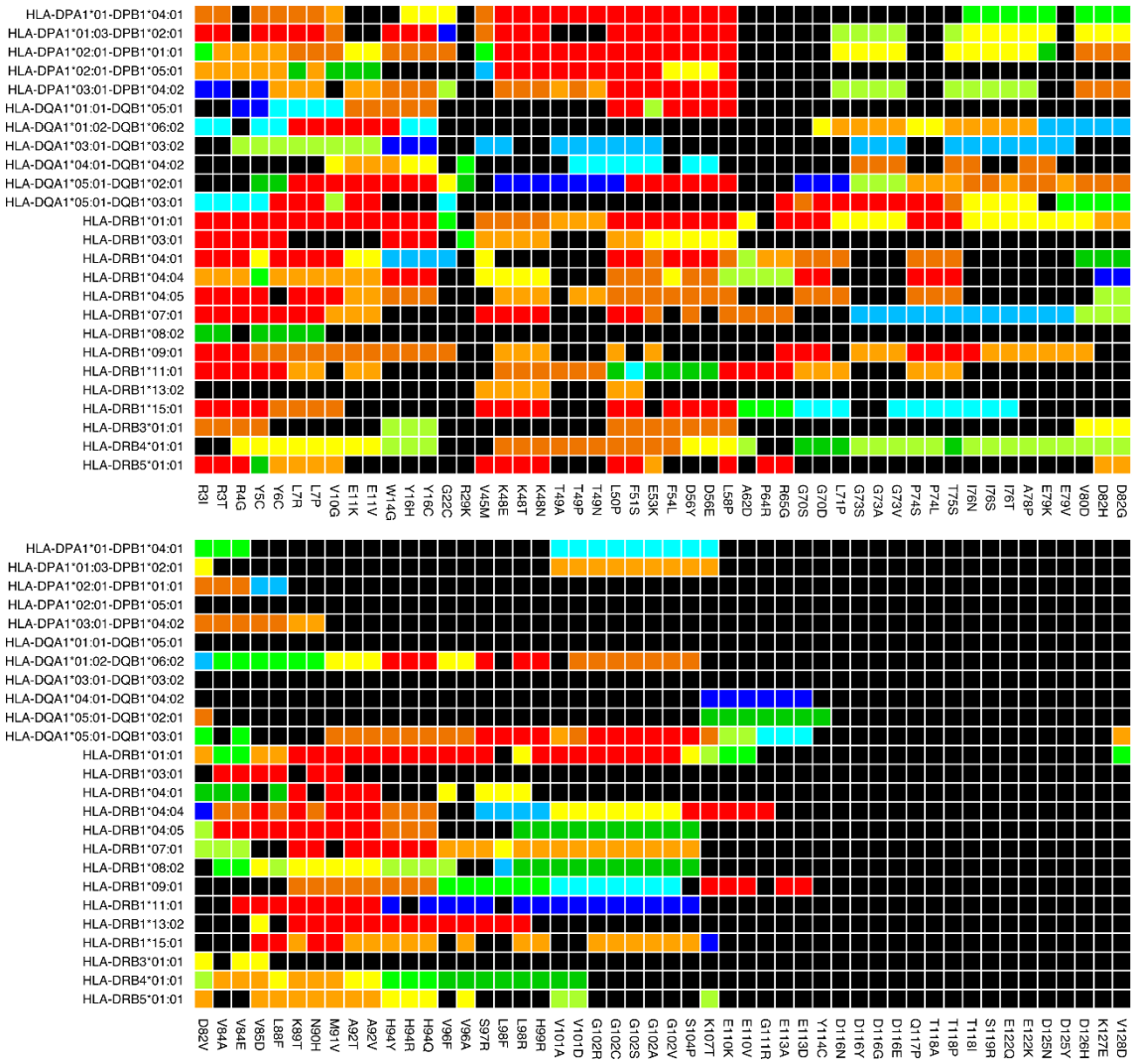
Zachary, A. A., & Leffell, M. S. (2016). *HLA Mismatching Strategies for Solid Organ Transplantation – A Balancing Act*. *7*(December), 1–14. https://doi.org/10.3389/fimmu.2016.00575

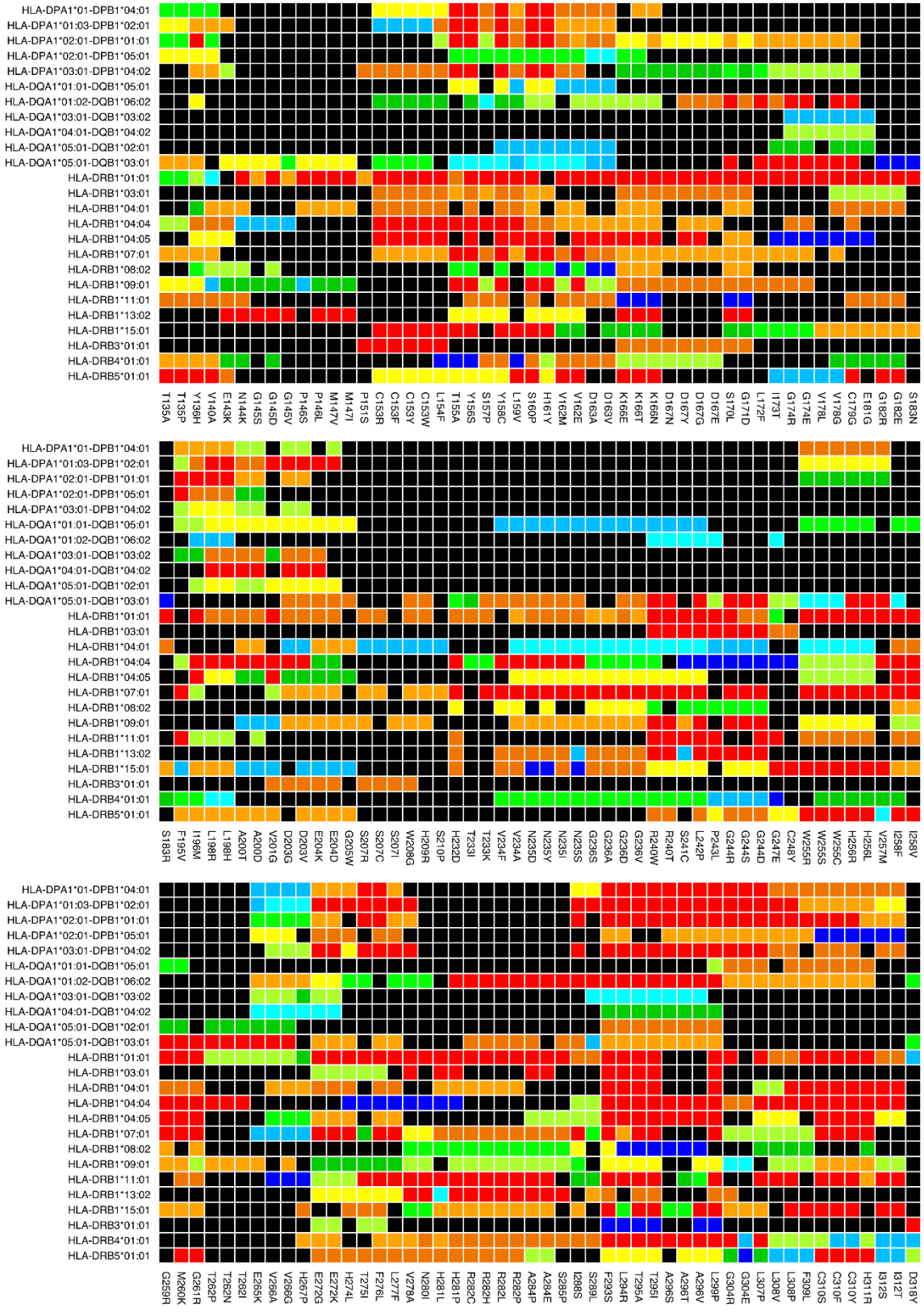
Zerbini, A., Pilli, M., Soliani, P., Ziegler, S., Pelosi, G., Orlandini, A., Cavallo, C., Uggeri, J., Scandroglio, R., Crafa, P., Spagnoli, G. C., Ferrari, C., & Missale, G. (2004). Ex vivo characterization of tumor-derived melanoma antigen encoding gene-specific CD8 + cells in patients with hepatocellular carcinoma. *Journal of Hepatology*, *40*(1), 102–109. https://doi.org/10.1016/S0168-8278(03)00484-7

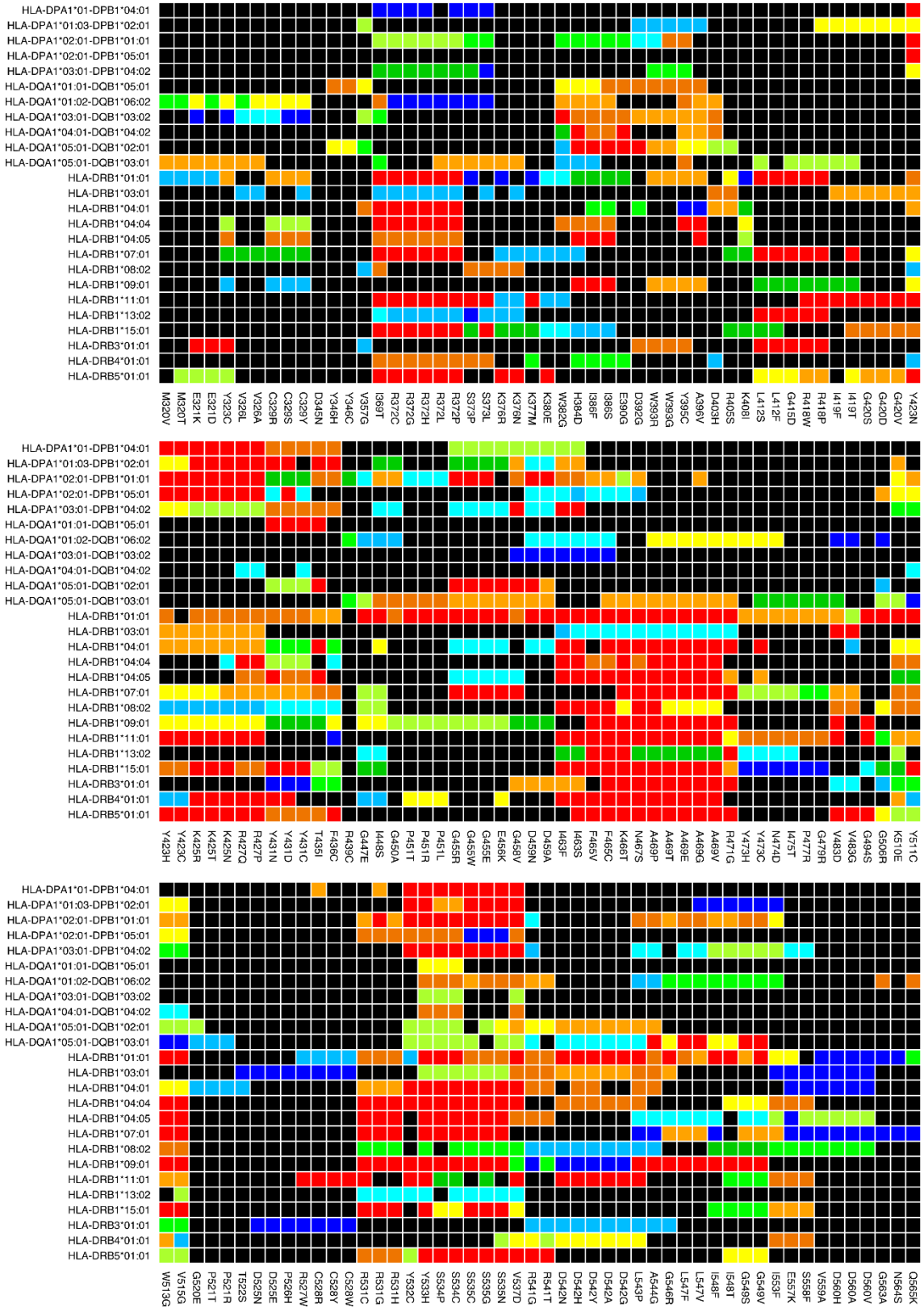
Zino, E., Frumento, G., Marktel, S., Sormani, M. P., Ficara, F., di Terlizzi, S., Parodi, A. M., Sergeant, R., Martinetti, M., Bontadini, A., Bonifazi, F., Lisini, D., Mazzi, B., Rossini, S., Servida, P., Ciceri, F., Bonini, C., Lanino, E., Bandini, G., … Fleischhauer, K. (2004). A T-cell epitope encoded by a subset of HLA-DPB1 alleles determines nonpermissive mismatches for hematologic stem cell transplantation. *Blood*, *103*(4), 1417–1424. https://doi.org/10.1182/blood-2003-04-1279

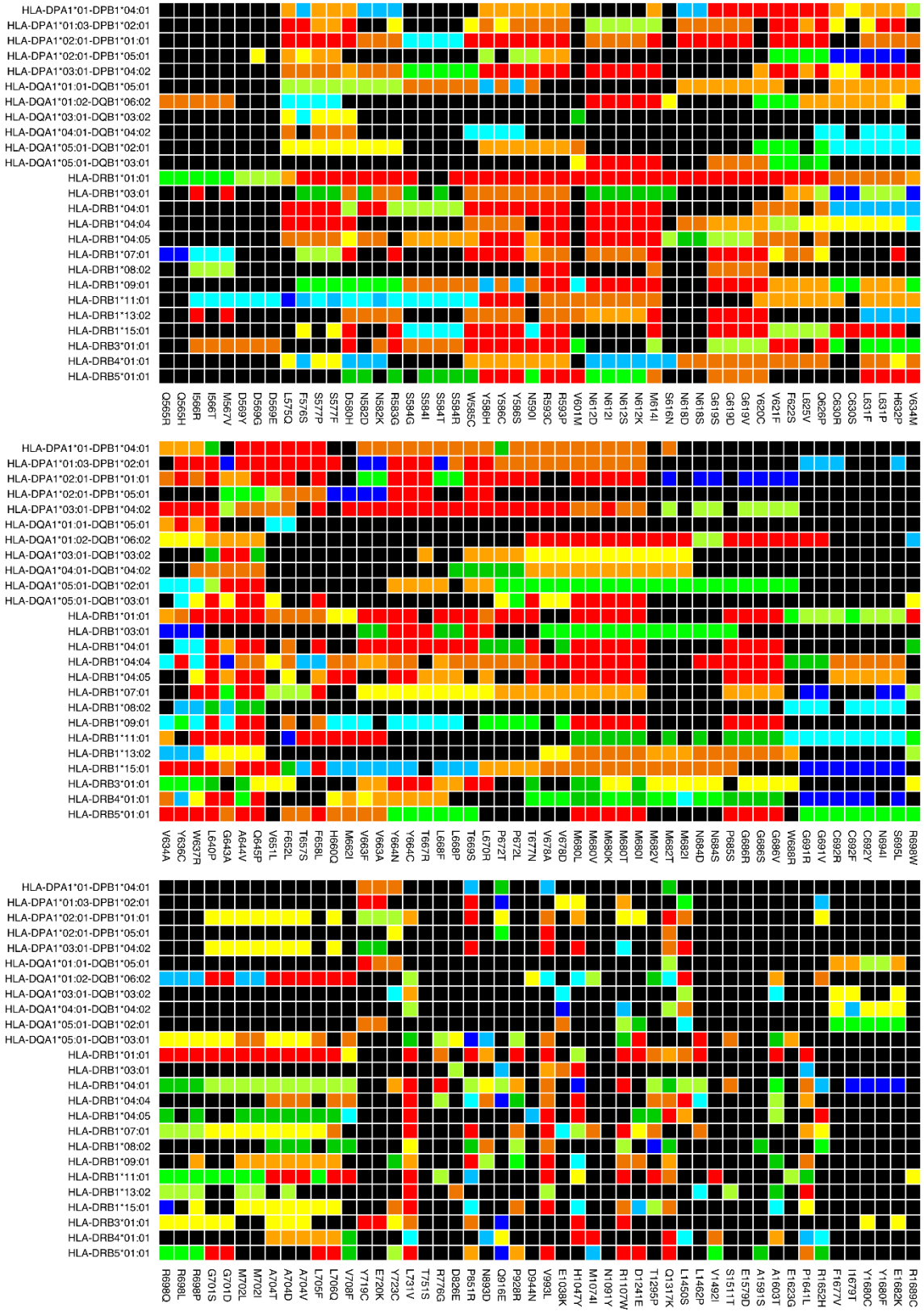
8 Appendices

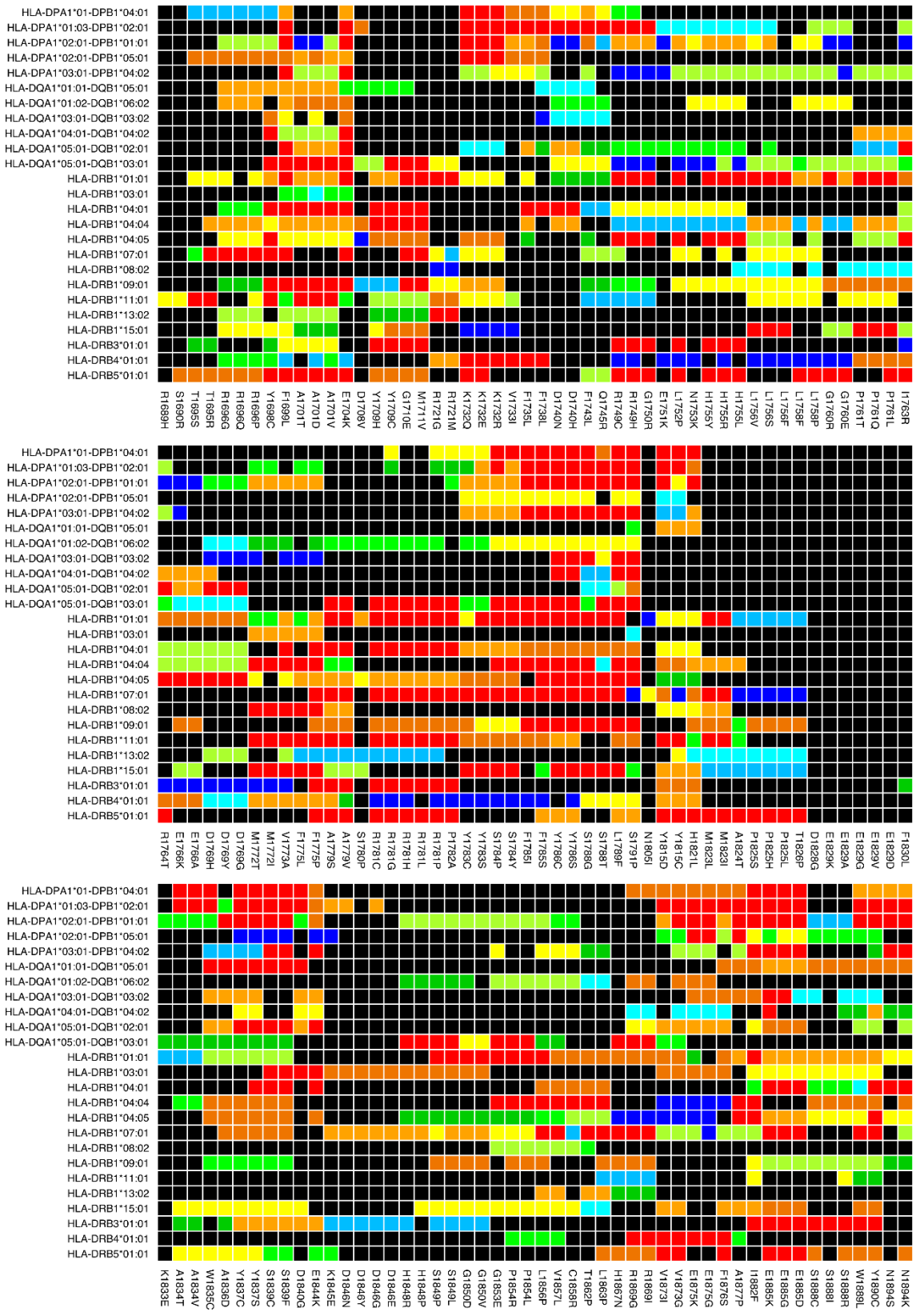
8.1 Appendix 1: Full heatmaps covering all FVIII missense mutations in the Factor VIII Gene (F8) Variant Database with and without proteome scanning

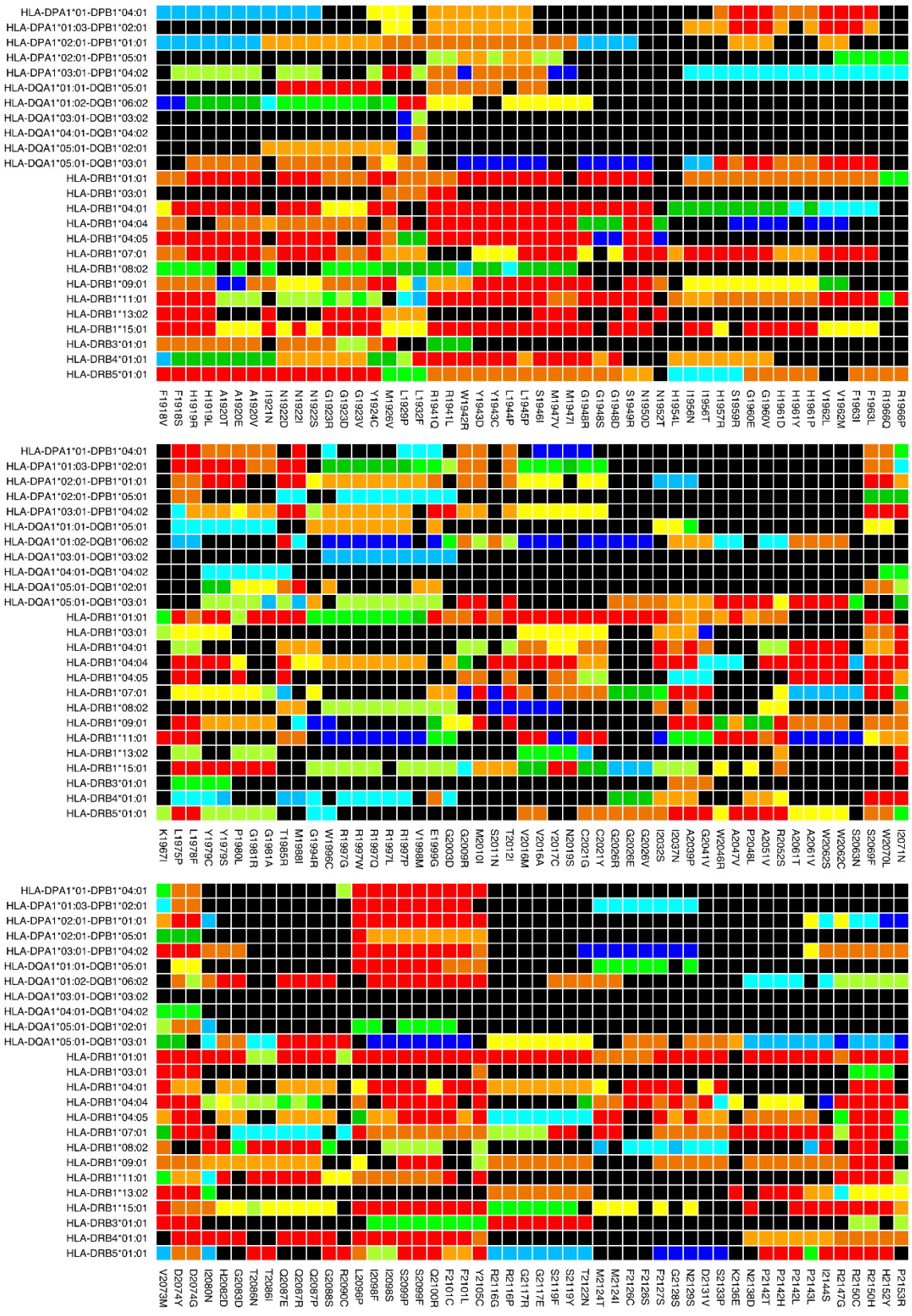



HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
T135A
T135P
Y136H
V140A
E143K
N144K
G145S
G145D
G145V
P146S
P146L
M147V
M147I
P151S
C153R
C153F
C153Y
C153W
L154F
T155A
Y155S
S157P
Y158C
L159V
S160P
H161Y
V162M
V162E
D163A
D163V
K166E
K166T
K166N
D167N
D167Y
D167G
D167E
S170L
G171D
L172F
I173T
G174R
G174E
V178L
V178G
C179G
E181G
G182R
G182E
S183N
HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
S183R
F195V
I196M
L198R
L198H
A200T
A200D
V201G
D203G
D203V
E204K
E204D
G205W
S207R
S207C
S207I
W208G
H209R
S210P
H232D
T233I
T233K
V234F
V234A
N235D
N235Y
N235I
N235S
G236S
G236A
G236D
G236V
R240W
R240T
S241C
L242P
P243L
G244R
G244S
G244D
G247E
C248Y
W255R
W255S
W255C
H256R
H256L
V257M
I258F
I258V
HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
G259R
M260K
G261R
T262P
T262N
T262I
E265K
V266A
V266G
H267P
E272G
E272K
H274L
T275I
F276L
L277F
V278A
N280I
H281L
H281P
R282C
R282H
R282L
R282P
A284P
A284E
S285P
I288S
S289L
F293S
L294R
T295A
T295I
A296S
A296T
A296V
L299P
G304R
G304E
L307P
L308V
L308P
F309L
C310S
C310F
C310Y
H311R
I312S
I312T
D318Y


HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
M320V
M320T
E321K
E321D
Y323C
V326L
V326A
C329R
C329S
C329Y
D345N
Y346H
Y346C
V357G
I369T
R372C
R372G
R372H
R372L
R372P
S373P
S373L
K376R
K376N
K377M
K380E
W382G
H384D
I386F
I386S
E390G
D392G
W393R
W393G
Y395C
A396V
D403H
R405S
K408I
L412S
L412F
G415D
R418W
R418P
I419F
I419T
G420S
G420D
G420V
Y423N
HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
Y423H
Y423C
K425R
K425T
K425N
R427Q
R427P
Y431N
Y431D
Y431C
T435I
F436C
R439C
G447E
I448S
G450A
P451T
P451R
P451L
G455R
G455W
G455E
E456K
G458V
D459N
D459A
I463F
I463S
F465V
F465C
K466T
N467S
A469P
A469T
A469E
A469G
A469V
R471G
Y473H
Y473C
N474D
I475T
P477R
G479R
V483D
V483G
G494S
G506R
K510E
Y511C
HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
W513G
V515G
G520E
P521T
P521R
T522S
D525N
D525E
P528H
R527W
C528R
C528Y
C528W
R531C
R531G
R531H
Y532C
Y533H
S534P
S534C
S535C
S535G
S535N
V537D
R541T
R541G
D542N
D542H
D542Y
D542A
D542G
L543P
A544G
G546R
L547F
L547V
I548F
I548T
G549S
G549V
I553F
E557K
S558F
V559A
D560H
D560A
D560V
G563A
N564S
Q565K


HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
Q565R
Q565H
I566R
I566T
M567V
D569Y
D569G
D569E
L575Q
F576S
S577P
S577F
D580H
N582D
N582K
R583G
S584G
S584I
S584T
S584R
W585C
Y586H
Y586C
Y586S
N590I
R593C
R593P
V601M
N612D
N612I
N612S
N612K
M614I
S616N
N618D
N618S
G619S
G619D
G619V
Y620C
V621F
F622S
L625V
Q626P
C630R
C630S
L631F
L631P
H632P
V634M
HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
V634A
Y636C
W637R
L640P
G643A
A644V
Q645P
V651L
F652L
T657S
F658L
H660Q
M662I
V663F
V663A
Y664N
Y664C
T667R
L668F
L668P
T669S
L670R
P672T
P672L
T677N
V678A
V678D
M680L
M680V
M680K
M680T
M680I
M682V
M682T
M682I
N684D
N684S
P685S
G686R
G686S
G686V
W688R
G691R
G691V
C692R
C692F
C692Y
N694I
S695L
R698W
HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
R698Q
R698L
R698P
G701S
G701D
M702L
M702I
A704T
A704D
A704V
A704V
L705F
L706Q
V708F
Y719C
E720K
Y723C
L731V
T751S
R776G
D826E
P851R
N893D
Q916E
P928R
D944N
V993L
E1038K
H1047Y
M1074I
N1091Y
R1107W
D1241E
T1295P
Q1317K
L1450S
L1462P
V1492I
S1511T
E1579D
A1591S
A1603T
E1623G
P1641L
R1652H
F1677V
I1679T
Y1680C
Y1680F
E1682K
R1689C


HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
R1689H
S1690R
T1695S
T1695R
R1696G
R1696Q
R1696P
Y1698C
F1699L
A1701T
A1701D
A1701V
E1704K
D1708V
Y1709H
Y1709C
G1710E
M1711V
R1721G
R1721M
K1732Q
K1732E
K1732R
V1733I
F1735L
F1738L
D1740N
D1740H
F1743L
Q1745R
R1749C
R1749H
G1750R
E1751K
L1752P
N1753K
H1755Y
H1755R
H1755L
L1756V
L1756S
L1756F
L1756F
L1758F
L1758P
G1760R
G1760E
P1761T
P1761Q
P1761L
I1763R
HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
R1764T
E1766K
E1766A
D1769H
D1769Y
D1769G
M1772T
M1772I
V1773A
F1775L
F1775P
A1779S
A1779V
S1780P
R1781C
R1781G
R1781H
R1781L
R1781P
P1782A
Y1783C
Y1783S
S1784P
S1784Y
F1785I
F1785S
Y1786C
Y1786S
S1788G
S1788T
L1789F
S1791P
N1805I
Y1815D
Y1815C
H1821L
M1823L
M1823I
A1824T
P1825S
P1825H
P1825L
T1826P
D1828G
E1829K
E1829G
E1829V
E1829D
E1830L
HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
K1833E
A1834T
A1834V
W1835C
A1836D
Y1837C
Y1837S
S1839C
S1839F
D1840G
E1844K
K1845E
D1846N
D1846Y
D1846G
D1846E
H1848R
H1848P
S1849P
S1849L
G1850D
G1850V
G1853E
P1854R
P1854L
L1856P
C1858R
T1862P
L1863P
H1867N
R1869G
R1869I
V1873I
V1873G
E1875K
E1875G
F1876S
A1877P
I1882F
E1885K
E1885G
E1885D
S1888C
S1888I
S1888R
W1889L
Y1890C
N1894S
N1894K


HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
F1918V
F1918S
H1919R
H1919L
A1920T
A1920E
A1920V
I1921N
N1922I
N1922D
N1922S
G1923R
G1923D
G1923V
Y1924C
M1926V
L1928P
L1932F
R1941Q
R1941L
W1942R
Y1943D
Y1943C
L1944P
L1945P
S1946I
M1947V
M1947I
G1948R
G1948S
G1948D
S1949R
N1950D
N1952T
H1954L
I1956N
I1956T
H1957R
S1959R
G1960E
G1960V
H1961D
H1961Y
H1961P
V1962L
V1962M
F1963I
F1963L
R1966Q
R1966P
HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
K1967I
L1975P
L1978F
Y1979C
Y1979S
P1980L
G1981R
G1981A
G1994R
M1988I
T1985R
W1996C
R1997G
R1997W
R1997Q
R1997L
R1997P
V1998M
E1999G
G2003D
G2009R
M2010I
S2011N
T2012I
V2016M
V2016A
Y2017C
N2019S
C2021G
C2021Y
G2026R
G2026E
G2026V
I2032S
I2037N
A2039P
G2041V
W2046R
A2047V
P2048L
A2051V
R2052S
A2061T
A2061V
W2062C
W2062S
S2063N
S2069F
W2070L
I2071N
HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
V2073M
D2074Y
D2074G
I2080N
H2082D
G2083D
T2086N
T2086I
Q2087E
Q2087R
Q2087P
G2088S
R2090C
L2096P
I2098F
I2098S
S2099P
S2099F
Q2100R
F2101C
F2101L
Y2105C
R2116G
R2116P
G2117R
G2117E
G2117F
S2119F
S2119Y
T2122N
M2124T
M2124I
F2126C
F2126S
F2127S
G2128S
N2129S
D2131V
S2133P
K2136E
N2138D
P2142T
P2142H
P2142L
P2143L
I2144S
R2147G
R2150C
R2150H
H2152Y
P2153R

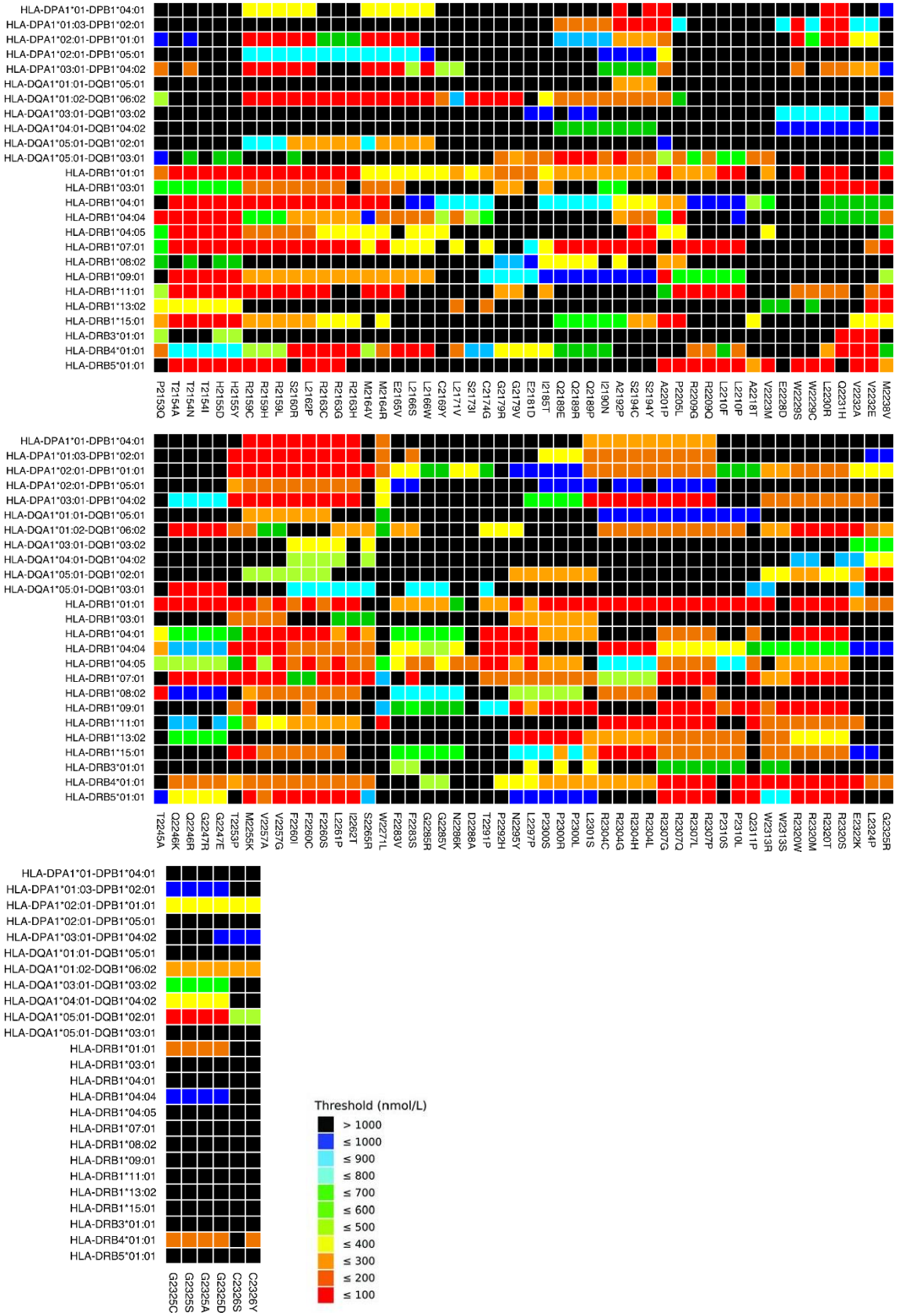
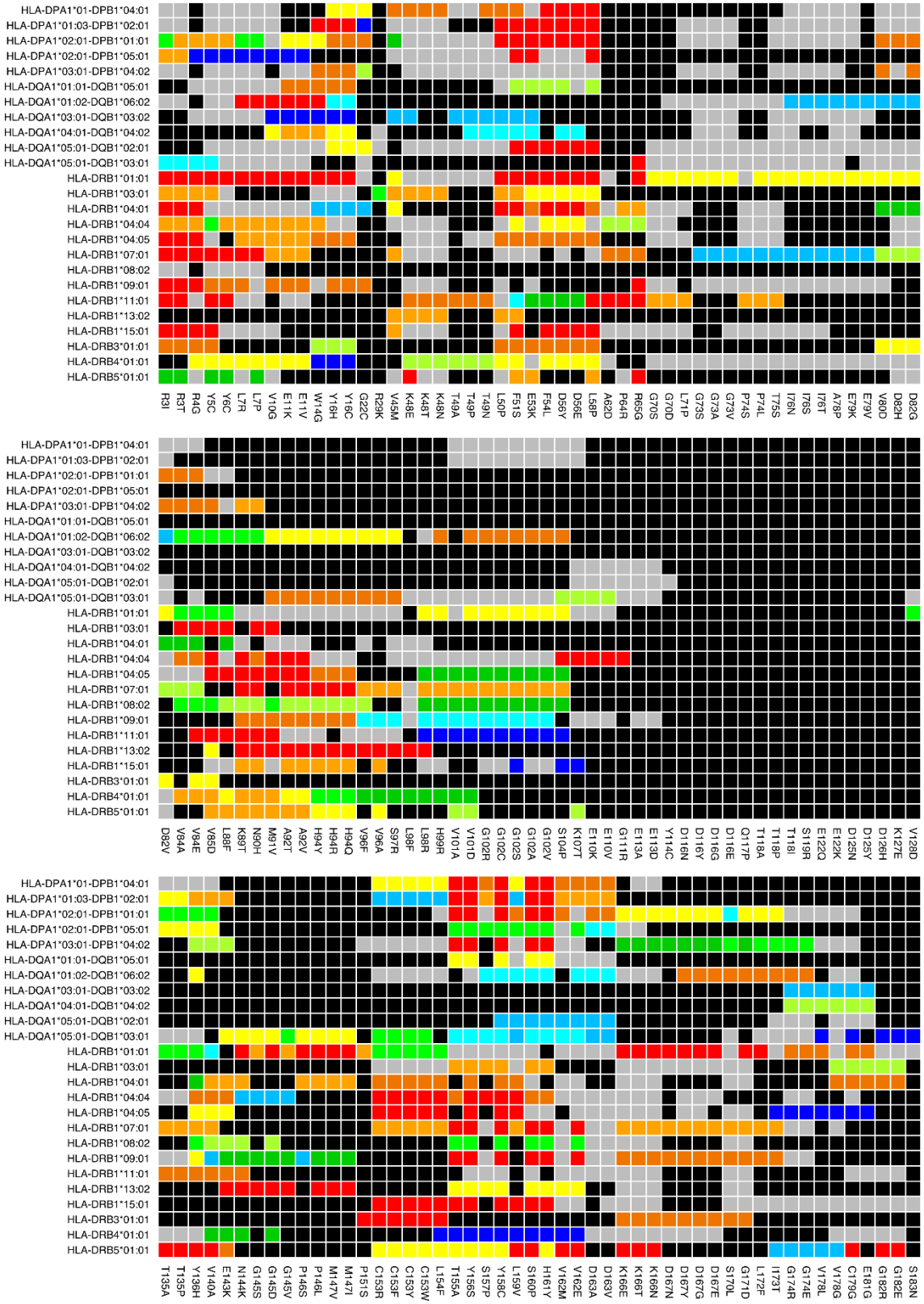
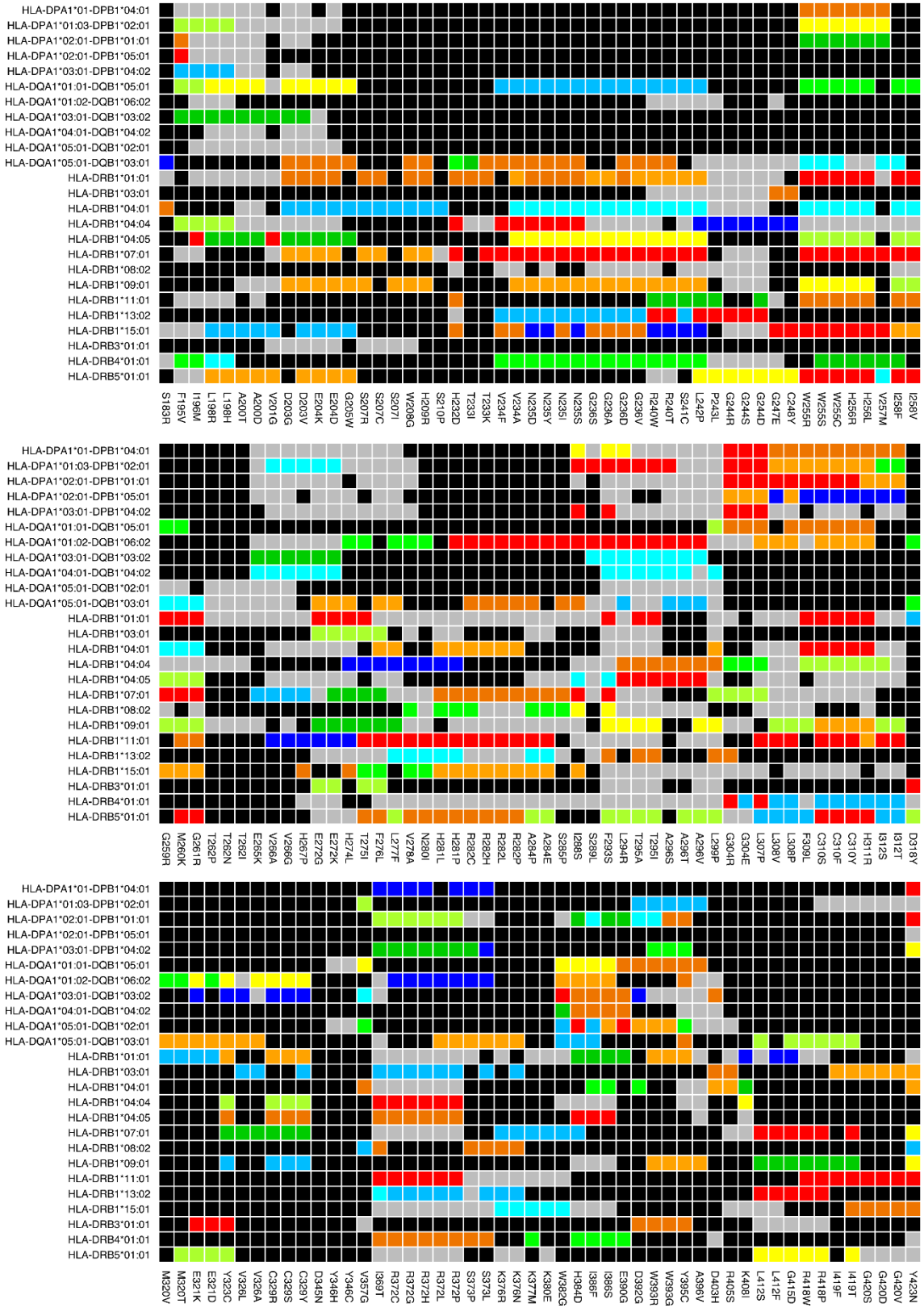
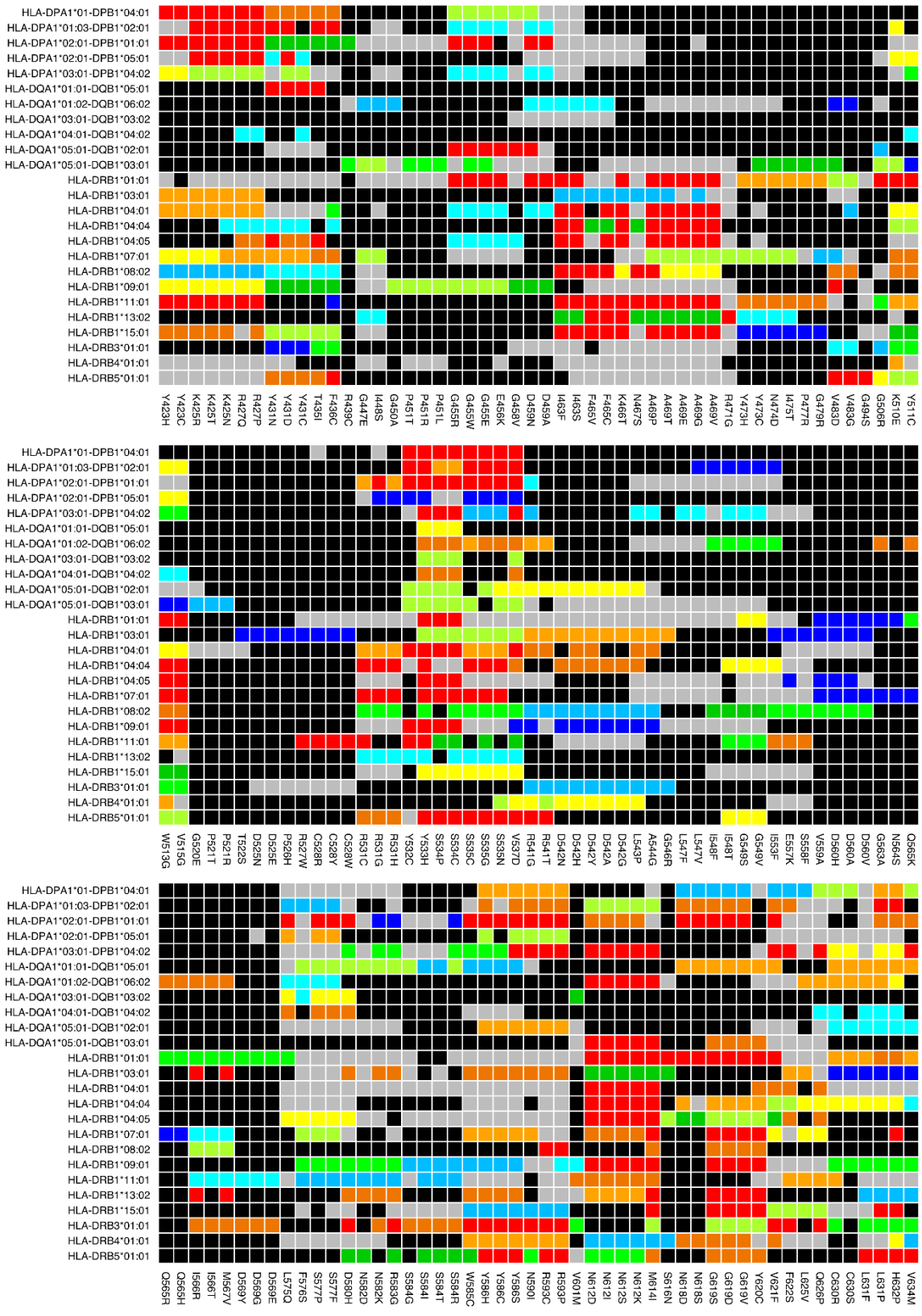
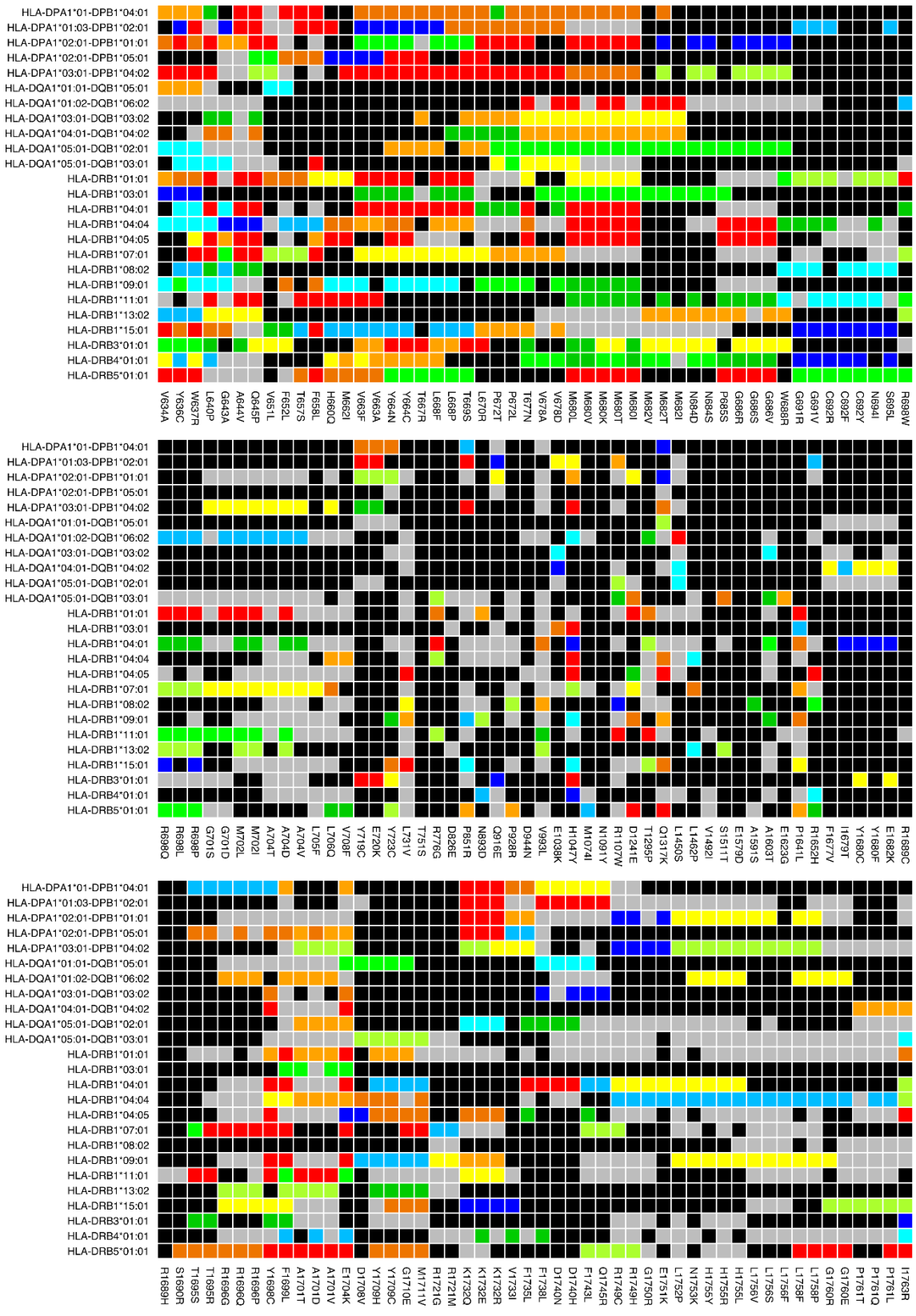


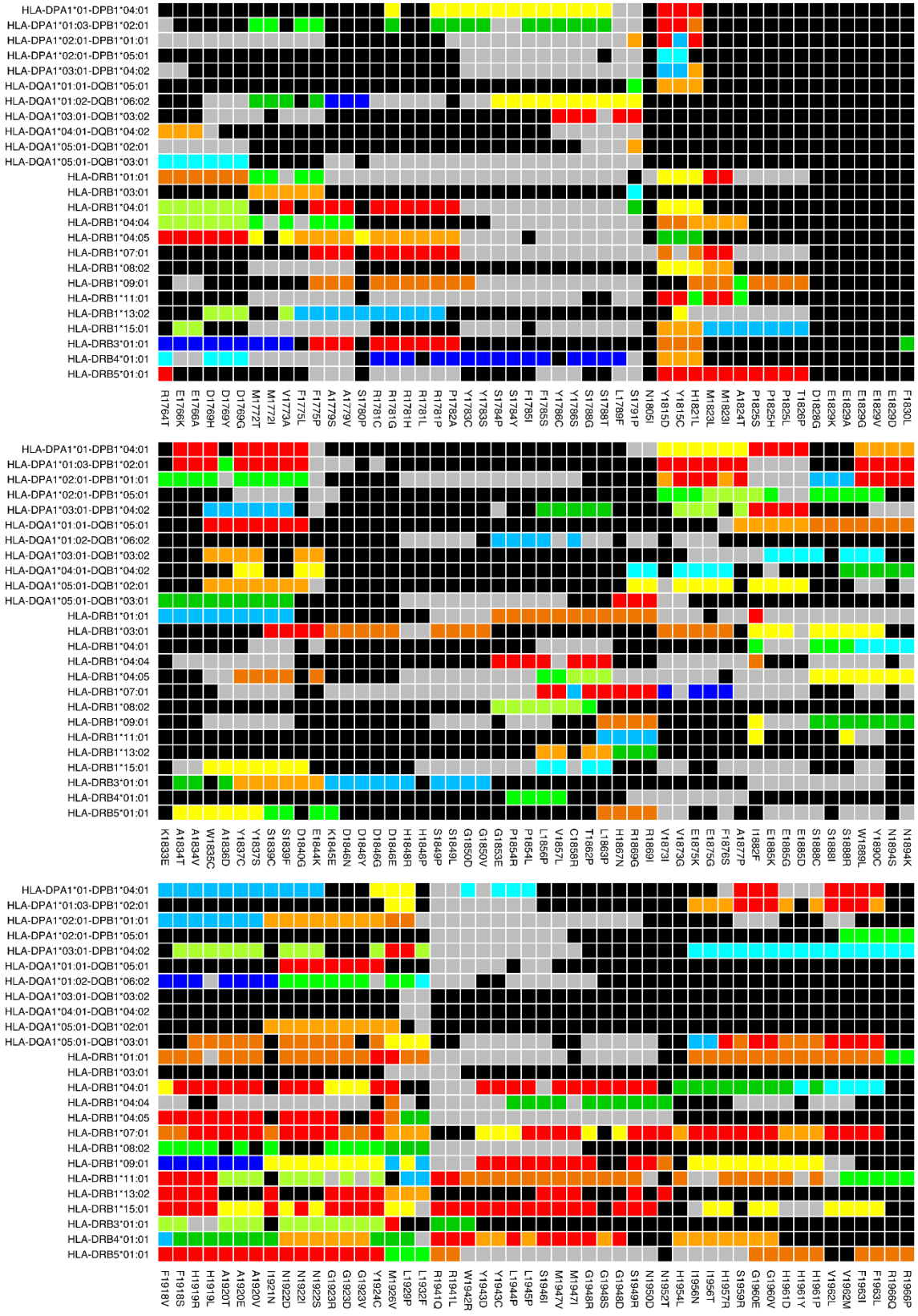
Figure A1 Heatmaps sowing all available missense mutations in the Factor VIII Gene (F8) Variant Database without proteome scanning

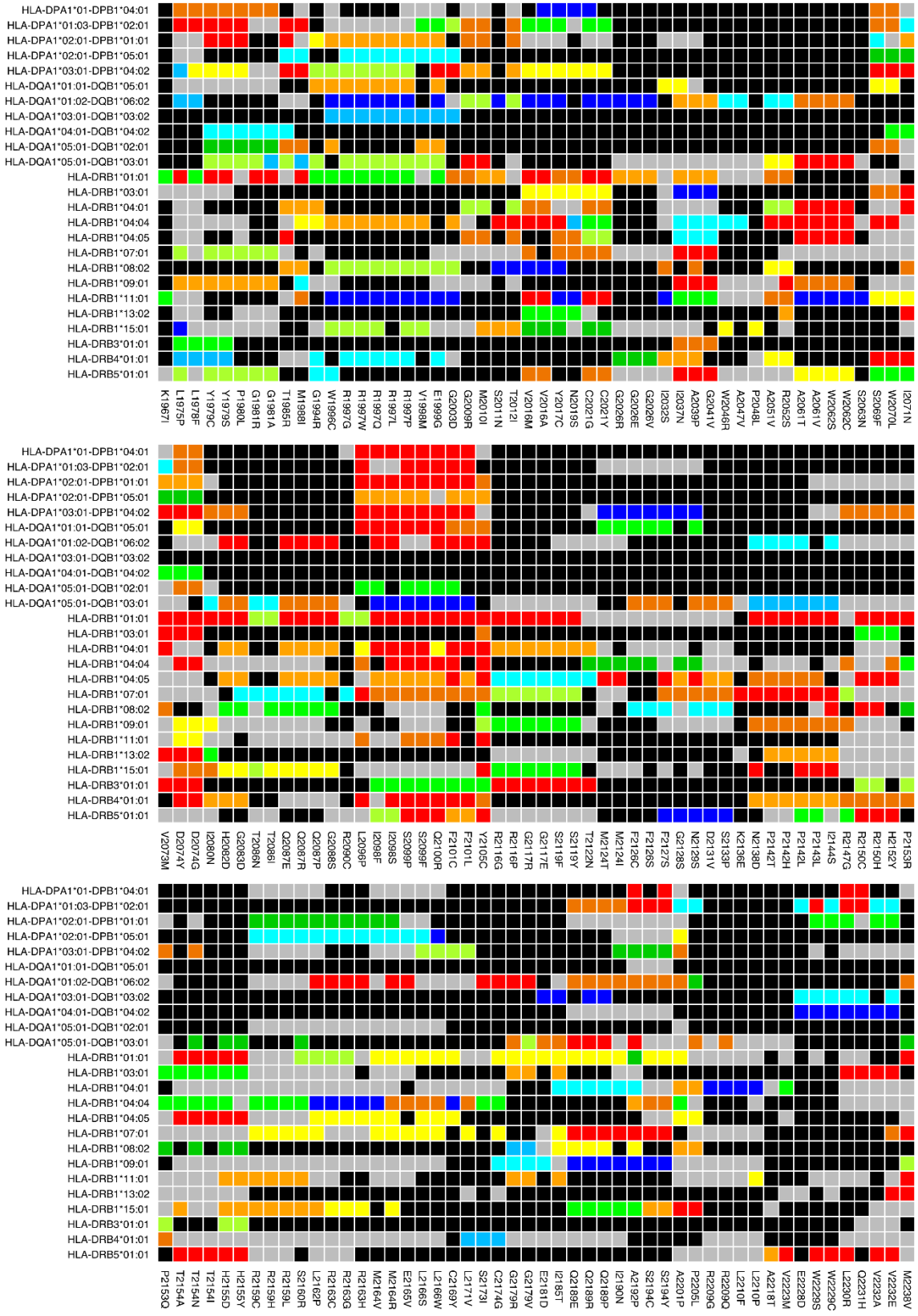

HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
R3I R3T R4G Y5C Y6C L7R L7P V10G E11K E11V W14G Y16H Y16C G22C R29K V45M K48E K48T K48N T49A T49P T49N L50P F51S E53K F54L D56Y D56E L58P A62D P64R R65G G70S G70D L71P G73S G73A G73V P74S P74L T75S I76N I76S I76T A78P E79K E79V V80D D82H D82G
HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
D82V V84A V84E V85D L88F K89T N90H M91V A92T A92V H94Y H94R H94Q V96F V96A S97R L98F L98R H99R V101A V101D G102R G102C G102S G102A G102V S104P K107T E110K E110V G111R E113A E113D Y114C D116N D116Y D116G D116E Q117P T118A T118P T118I S119R E122Q E122K D125N D125Y D126H K127E V128D
HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
T135A T135P Y136H V140A E143K N144K G145S G145D G145V P146S P146L M147V M147I P151S C153R C153F C153Y C153W L154F T155A Y156S S157P Y158C L159V S160P V162M V162E D163A D163V K166E K166T K166N D167N D167Y D167G D167E S170L G171D L172F I173T G174R G174E V178L V178G C179G E181G G182R G182E S183N


HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
S183R
F195V
I196M
L198R
L198H
A200T
A200D
V201G
D203G
D203V
E204K
E204D
G205W
S207R
S207C
S207I
W208G
H209R
S210P
H232D
T233I
T233K
V234F
V234A
N235D
N235Y
N235I
N235S
G236S
G236A
G236D
G236V
R240W
R240T
S241C
L242P
P243L
G244R
G244S
G244D
G247E
C248Y
W255R
W255S
W255C
H256R
H256L
V257M
I258F
I258V
HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
G259R
M260K
G261R
T262P
T262N
T262I
E265K
V266A
V266G
H267P
E272G
E272K
H274L
T275I
F276L
L277F
V278A
N280I
H281L
H281P
R282C
R282H
R282L
R282P
A284P
A284E
S285P
I288S
S289L
F293S
L294R
T295A
T295I
A296S
A296T
A296V
L299P
G304R
G304E
L307P
L308V
L308P
F309L
C310S
C310F
C310Y
H311R
I312S
I312T
D316Y
HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
M320V
M320T
E321K
E321D
Y323C
V326L
V326A
C329R
C329S
C329Y
D345N
Y346H
Y346C
V357G
I369T
R372C
R372G
R372H
R372L
R372P
S373P
S373L
K376R
K376N
K377M
K380E
W382G
H384D
I386F
I386S
E390G
D392G
W393R
W393G
Y395C
A396V
D403H
R405S
K408I
L412S
L412F
G415D
R418W
R418P
I419F
I419T
G420S
G420D
G420V
Y423N


HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
Y423H
Y423C
K425R
K425T
K425N
R427Q
R427P
Y431N
Y431D
Y431C
T435I
F436C
R439C
G447E
I448S
G450A
P451T
P451R
P451L
G455R
G455W
G455E
E456K
G458V
G458N
D459A
D459N
I463F
I463S
F465V
F465C
K466T
N467S
A469P
A469T
A469E
A469G
A469V
R471G
Y473H
Y473C
N474D
I475T
P477T
P477R
G479R
V483D
V483G
G494S
G506R
K510E
Y511C
HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
W513G
V515G
G520E
P521T
P521R
T522S
D525N
D525E
P526H
R527W
C528R
C528Y
C528W
R531C
R531G
R531H
Y532C
Y533H
S534P
S534C
S535C
S535G
S535N
V537D
R541G
R541T
D542N
D542H
D542Y
D542A
D542G
L543P
A544G
G546R
L547F
L547V
I548F
I548T
G549S
G549V
I553F
E557K
S558F
V559A
D560H
D560A
D560V
G563A
N564S
Q565K
HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
Q565R
Q565H
I566R
I566T
M567V
D569Y
D569G
D569E
L575Q
F576S
S577P
S577F
D580H
N582D
N582K
R583G
S584G
S584I
S584T
S584R
W585C
Y586H
Y586C
Y586S
N590I
R593C
R593P
V601M
N612D
N612I
N612S
N612K
M614I
S616N
N618D
N618S
G619S
G619D
G619V
Y620C
V621F
F622S
L625V
Q626P
C630R
C630S
L631F
L631P
H632P
V634M


HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
V634A
Y636C
W637R
L640P
G643A
A644V
Q645P
V651L
F652L
T657S
F658L
H660Q
M662I
V663F
V663A
Y664N
Y664C
T667R
L668F
L668P
T669S
L670R
P672T
P672L
T677N
V678A
V678D
M680L
M680V
M680K
M680T
M680I
M682V
M682T
M682I
N684D
N684S
P685S
G686R
G686S
G686V
W688R
G691R
G691V
C692R
C692F
C692Y
N694I
S695L
R698W
HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
R698Q
R698L
R698P
G701S
G701D
M702L
M702I
A704T
A704D
A704V
L705F
L706Q
V708F
Y719C
E720K
Y723C
L731V
T751S
R776G
D826E
P851R
N893D
Q916E
P928R
D944N
V993L
E1038K
H1047Y
M1074I
N1091Y
R1107W
D1241E
T1295P
Q1317K
L1450S
L1462P
V1492I
S1511T
E1579D
A1591S
A1603T
E1623G
P1641L
R1652H
I1679T
F1677V
Y1680C
Y1680F
E1682K
R1689C
HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
R1689H
S1690R
T1695S
T1695R
R1696G
R1696Q
R1696P
Y1698C
F1699L
A1701T
A1701D
A1701V
E1704K
D1708Y
Y1709H
Y1709C
G1710E
M1711V
R1721G
R1721M
K1732Q
K1732E
K1732R
V1733I
F1735L
F1738L
D1740N
D1740H
F1743L
Q1745R
R1749C
R1749H
G1750R
E1751K
L1752P
N1753K
H1755Y
H1755R
H1755L
L1756V
L1756S
L1756F
L1758F
L1758P
G1760R
G1760E
P1761T
P1761Q
P1761L
I1763R


HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
R1764T
E1766K
E1766A
D1769H
D1769Y
D1769G
M1772T
M1772I
V1773A
F1775L
F1775P
A1779S
A1779V
S1780P
R1781C
R1781G
R1781H
R1781L
R1781P
P1782A
Y1783C
Y1783S
S1784P
S1784Y
F1785I
F1785S
Y1786C
Y1786S
S1788G
S1788T
L1789F
S1791P
N1805I
Y1815D
Y1815C
H1821L
M1823L
M1823I
A1824T
P1825S
P1825H
P1825L
T1826P
D1828G
E1829K
E1829A
E1829G
E1829V
E1829D
F1830L
HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
K1833E
A1834T
A1834V
W1835C
A1836D
Y1837C
Y1837S
S1839C
S1839F
D1840G
E1844K
K1845E
D1846N
D1846Y
D1846G
D1846E
H1848R
H1848P
S1849P
S1849L
G1850D
G1850V
G1853E
P1854R
P1854L
L1856P
V1857L
C1858R
T1862P
L1869P
H1867N
R1869G
R1869I
V1873I
V1873G
E1875K
E1875G
F1876S
A1877P
I1882F
E1885K
E1885G
E1885D
S1888C
S1888I
S1888R
W1889L
Y1890C
N1894S
N1894K
HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
F1918V
F1918S
H1919R
H1919L
A1920T
A1920E
A1920V
I1921N
N1922D
N1922I
N1922S
G1923R
G1923D
G1923V
Y1924C
M1926V
L1929P
L1932F
R1941Q
R1941L
W1942R
Y1943D
Y1943C
L1944P
L1945P
S1946I
M1947V
M1947I
G1948R
G1948S
G1948D
S1949R
N1950D
N1952T
H1954L
I1956N
I1956T
H1957R
S1959R
G1960E
G1960V
H1961D
H1961Y
H1961P
Y1962L
Y1962M
F1963I
F1963L
R1966Q
R1966P


HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
K1967I
L1975P
L1978F
Y1979C
Y1979S
P1980L
G1981R
G1981A
T1985R
M1988I
G1994R
W1996C
R1997G
R1997W
R1997Q
R1997L
R1997P
V1998M
E1999G
G2003D
G2009R
M2010I
S2011N
T2012I
V2016M
V2016A
Y2017C
N2019S
C2021G
C2021Y
G2026R
G2026E
G2026V
I2032S
I2037N
A2039P
G2041V
W2046R
A2047V
P2048L
A2051V
R2052S
A2061T
A2061V
W2062S
W2062C
S2063N
S2069F
W2070L
I2071N
HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
V2073M
D2074Y
D2074G
I2080N
H2082D
G2083D
T2086N
T2086I
Q2087E
Q2087R
Q2087P
G2088S
R2090C
L2096P
I2098F
I2098S
S2099P
S2099F
Q2100R
F2101C
F2101L
Y2105C
R2116G
R2116P
G2117R
G2117E
S2119F
S2119Y
T2122N
M2124T
M2124I
F2126C
F2126S
F2127S
G2128S
N2129S
D2131V
S2133P
K2136E
N2138D
P2142T
P2142H
P2142L
P2143L
21144S
R2147G
R2150C
R2150H
H2152Y
P2153R
HLA-DPA1*01-DPB1*04:01
HLA-DPA1*01:03-DPB1*02:01
HLA-DPA1*02:01-DPB1*01:01
HLA-DPA1*02:01-DPB1*05:01
HLA-DPA1*03:01-DPB1*04:02
HLA-DQA1*01:01-DQB1*05:01
HLA-DQA1*01:02-DQB1*06:02
HLA-DQA1*03:01-DQB1*03:02
HLA-DQA1*04:01-DQB1*04:02
HLA-DQA1*05:01-DQB1*02:01
HLA-DQA1*05:01-DQB1*03:01
HLA-DRB1*01:01
HLA-DRB1*03:01
HLA-DRB1*04:01
HLA-DRB1*04:04
HLA-DRB1*04:05
HLA-DRB1*07:01
HLA-DRB1*08:02
HLA-DRB1*09:01
HLA-DRB1*11:01
HLA-DRB1*13:02
HLA-DRB1*15:01
HLA-DRB3*01:01
HLA-DRB4*01:01
HLA-DRB5*01:01
P2153Q
T2154A
T2154N
T2154I
H2155D
H2155Y
R2159C
R2159H
R2159L
S2160R
L2162P
R2163C
R2163G
R2163H
M2164V
M2164R
E2165V
L2166S
L2166W
C2169Y
L2171V
S2173I
C2174G
G2179R
G2179V
E2181D
I2185T
Q2189E
Q2189R
Q2189P
I2190N
A2192P
S2194C
S2194Y
A2201P
P2205L
R2209G
R2209Q
L2210F
L2210P
A2218T
V2223M
E2228D
W2229S
W2229C
L2230R
Q2231H
V2232A
V2232E
M2238V

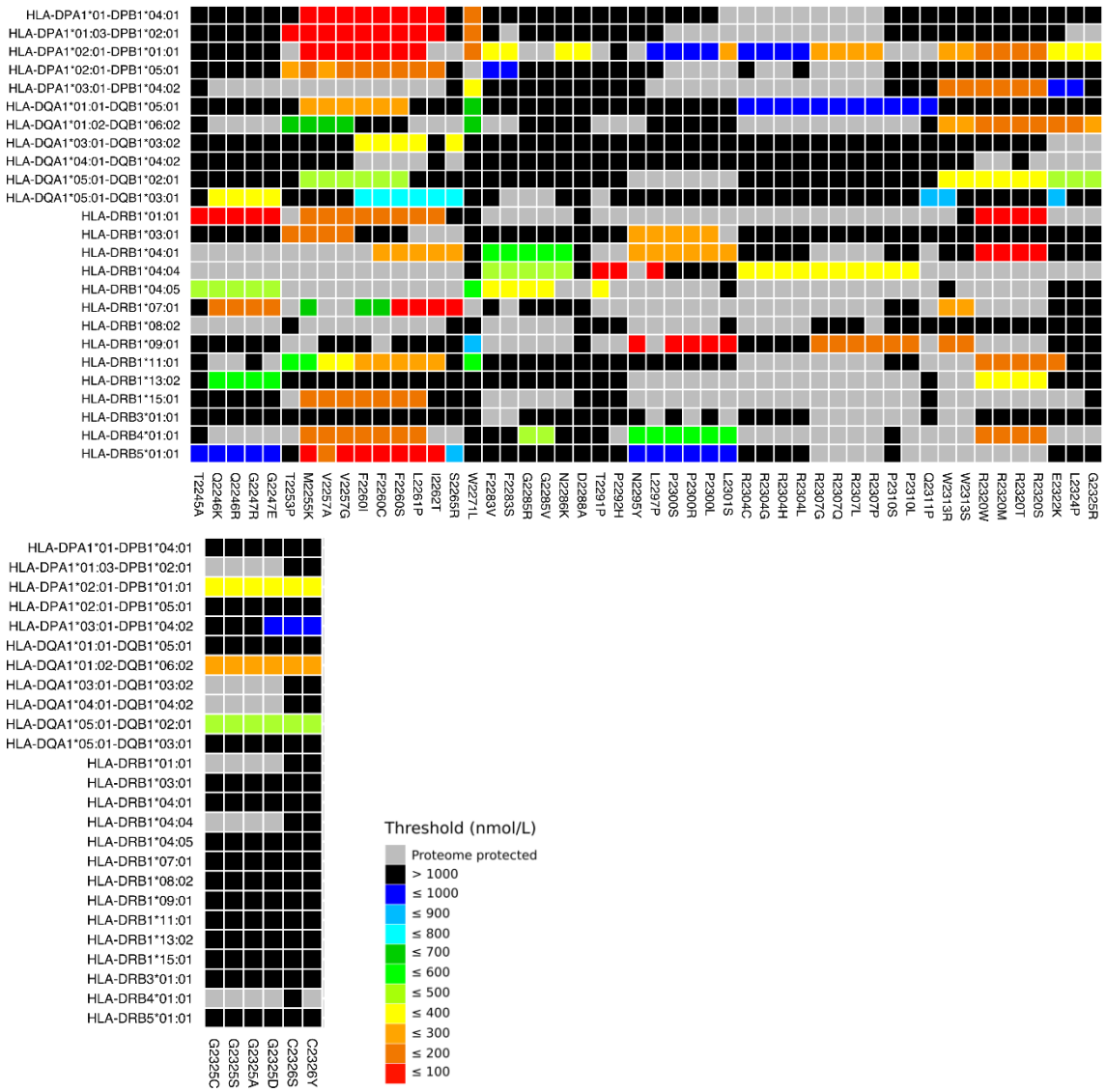


Figure A2 Heatmaps sowing all available missense mutations in the Factor VIII Gene (F8) Variant Database with proteome scanning